TLL Working Paper
December 2013

Development of a Framework to Classify MOOC Discussion Forum Posts: Methodology and Challenges

Glenda S. Stump[1], Jennifer DeBoer[1], Jonathan Whittinghill[2], Lori Breslow[1]

[1] Massachusetts Institute of Technology [2] Harvard Graduate School of Education

Abstract

The purpose of this paper is to describe the methodology used to confront one of the challenges associated with analyzing discussion forum data from the inaugural edX MOOC, "Circuits and Electronics." We detail the development and testing of a framework to classify large numbers of posts into a manageable number of categories so that further analysis can be conducted in targeted areas of interest. We discuss challenges that arose during implementation of the framework as well as how we resolved them. The resultant coding framework enables us to provide an empirical account of how students actually use the discussion forum in a MOOC--information that may also assist developers as they strive to enhance positive outcomes from students' use of this medium. Additionally, this framework can provide the foundation for the future utilization of natural language processing to code discussion forum data.

Massive Open Online Course (MOOC) discussion forums provide educational researchers with extraordinarily large quantities of rich data for analysis. The purpose of this paper is to describe the methodology used to confront the challenges of analyzing discussion forum data from the inaugural edX MOOC, "Circuits and Electronics" (also known as 6.002x), which ran from March through June of 2012. By November 2012, students had initiated 12,696 threads, or separate conversations, on the discussion forum; those threads garnered over 96,696 individual posts. From IP addresses recorded during each student interaction with the site, we identified that 194 countries were represented, and when asked for 'language' upon registration, students from those countries listed over 20 languages. These figures illustrate that discussion forum data generated in this course was indeed massive and potentially originated from a diverse population of students.

Within a MOOC discussion forum, students can voluntarily participate in collaborative exchanges with their peers. As a constructivist learning strategy, student collaboration provides an opportunity not only for exchange of ideas or critiques, but also for negotiation of meaning and co-construction of knowledge (Hull & Saxon, 2009; Jeong & Chi, 1997). Collaboration has been associated with the use of adaptive learning strategies such as knowledge-building (Husman, Lynch, Hilpert, Duggan, 2007; Salovaara, 2005; Schellens & Valcke, 2006; Shell et al., 2005) and improved learning outcomes (e.g., Smith et al., 2009; Springer, Stann, & Donovan, 1999). However, there are also null or negative benefits to engaging in collaborative activities with others (Barron, 2003; Salomon & Globerson, 1989), and further investigation into the nature of productive collaboration has been a consistent goal for educational researchers.

Regardless of their positive or negative outcomes, collaborative discussions provide a rich medium through which we can gain insight into cognitive processes of the participants. Analysis of collaborative discussions provides not only information related to individuals' level of content knowledge about a particular topic (e.g., Schrire, 2006; Weinberger & Fischer, 2006), but also about their learning strategies (Salovaara, 2005) or their social and communication skills (e.g., Järvelä & Häkkinen, 2002; Rourke et al., 1999). Exchanges that include co-constructing knowledge between learners, asking for help, or scaffolding another's understanding of difficult subjects require collaborators to utilize social and communication skills in conjunction with their content knowledge. Effective integration of these skills is essential to students' successful knowledge-building via collaboration.

Previously, researchers had one avenue to the analysis of collaborative learning discussions—digital or video recording of the interaction and subsequent transcription of the dialogue, a time-intensive task. The advent of discussion forums in online courses provided an already-transcribed account of learner interactions and eliminated this time-consuming task of transcription. An added advantage to use of this type of data is that students are not visually reminded that their dialogue will be dissected and analyzed when posting on a discussion forum as they are when speaking into a recorder or video camera. Studies of student interaction in traditional online discussion forums have added to our understanding of students' knowledge-building behaviors through collaborative discussion (e.g., Kortemeyer, 2006; Song & McNary, 2011) as well as peer tutoring (DeSmet, Van Keer, & Valke, 2006). The more recent development of MOOCs has provided yet another boon to analysis of learner interaction—larger amounts of data that are retrievable in text format from an extremely diverse group of learners (e.g., Breslow et al., 2013; DeBoer, Stump, Seaton, & Breslow, 2013; Kizilcec, Piech, & Schneider, 2013). However, analysis of this new richer source of data also introduces some challenging issues for educational researchers. Among those challenges is the daunting task of classifying or developing a description of thousands upon thousands of student posts. This first step is essential before undertaking further analysis to derive valid inferences and develop meaningful conclusions from the data.

Our specific research question for this work was, 'What is the topic of students' posts and the role assumed by posters on an open discussion forum?' Answering this question will inform our future work which focuses on determining the characteristics of productive dialogue between students in this space, as well as on exploring the relationships between types of dialogue utilized among students (e.g., content-related, social-related, technology-related), their role in that dialogue (e.g., help-seeker, help-giver, expert, novice), and their achievement or persistence in the course. In the sections that follow, we review relevant work in this area, describe development of a framework for analysis, and explain our approach to navigating the challenges we encountered when pursuing answers to these questions.

Relevant Work in Computer Supported Collaborative Learning

Collaborative exchanges on MOOC discussion forums fall into the category of computer supported collaborative learning (CSCL). Kirschner and Erkens (2013) named three main elements that contribute to students learning collaboratively in a computer-mediated

environment. The first is a pedagogical element, which refers to the means used to help students achieve the intended learning outcomes of instruction. For example, an instructor might use scripts that specify activities to facilitate collaborative dialogue among students (e.g., Fischer, Kollar, Stegmann, & Wecker, 2013) or representational tools that assist students to complete distinct tasks as they learn (e.g., Slof, Erkens, Kirschner, Jaspers, & Janssen, 2010). The second element is a social element, which refers to support for formation of functional groups and effective communication among group members. For example, an instructor might employ tools such as Team Maker (Layton, Loughry, Ohland, & Ricco, 2010) for group creation or Comprehensive Assessment of Team Member Effectiveness (CATME; Ohland, et al., 2012) to encourage personal reflection on contributions by group members. The third element is the technological element, which refers to functionality and restrictions of the CSCL environment. Pedagogical and social support via the tools presented as examples above would not be possible without adequate technological capability.

The 6.002x platform provided the technological capability for multiple simultaneous discussion threads that were tagged by topic and could be searched by key words. In its first iteration, the forum provided some pedagogical assistance to direct students' collaboration, in that they could search for tagged posts related to a topic of interest. Students were encouraged to do this prior to posting a question, thus allowing them to find a group of peers that may have shared their question and were still engaging in active discussion about it. In addition to asking or answering a question on the forum, students had the option to comment on questions and replies. Their comments often contained clarifications to questions or replies, which served to further direct the subsequent posts. Students were strongly encouraged to use the forum to find help with understanding course content or using course technology, but they could choose whether or not they wanted to use it.

Kirschner and Erkens (2013) further proposed a theoretical framework to guide CSCL research and development. They outlined three important dimensions of CSCL and related subcategories that should be addressed by research: (1) aspect of learning—the cognitive, social, and/or motivational outcomes of collaborative exchanges, (2) unit of learning—the benefit realized by an individual, group, team, or community from learning in collaborative endeavors, and (3) pedagogical measures—the interactive, representational, or guiding strategies used to facilitate positive outcomes for students as they collaborate. At present, CSCL research has focused on some, but not all, of these areas (e.g., Fransen, Weinberger, & Kirschner, 2013; Järvelä & Hadwin, 2013; Schellens & Valcke, 2006; Shell et al., 2005; Slof et al., 2010; Weinberger, Ertl, Fischer, & Mandl, 2005). Although our more immediate question of classifying the type of student posts and the role of the poster in the 6.002x discussion forum was descriptive, it contributes to our future work, which is focused specifically on the cognitive and social aspects of learning from the 6.002x discussion forum for individuals.

When taking a more fine-grained view of CSCL, the literature provides multiple examples of frameworks used for content analysis of discussion forum posts (see De Wever, Schellens, Valcke, & Van Keer, 2006 for a review). These frameworks have been used to examine

participants' interactivity, cognitive knowledge, metacognitive knowledge, critical thinking skills, social presence, perspective taking, and learning strategies. Although we believe these frameworks are extremely valuable to delve more deeply into students' learning behaviors and cognitions, we found that the open nature of the 6.002x discussion forum tended to produce a significant number of posts that were not strongly related to any of those constructs. In the MOOC space, it becomes necessary to precede more in-depth examinations of student posts with an initial coding or classification schema, to parse the sample prior to further analysis. We found no previous work to serve this purpose; thus, our work can fill a current void in MOOC-related CSCL research as well as provide a foundation for the future use of natural language processing in classification of discussion forum data.

Method

Data

Our data sources for this analysis were student posts retrieved from a database that saved every student post along with any subsequent edits on the discussion forum. Access to the discussion forum was prominently displayed as a tab on the course site, and a section entitled 'How to use the question / answer forum' was available on the course information page. After clicking on the discussion forum tab, students were directed to a page on which initial posts, known as 'threads,' were visible. This page could be sorted by date so that more recent threads appeared first; by activity so that threads with the most recent responses appeared first; by answers so that threads with the greatest number of responses appeared first; or by votes so that threads with the greatest number of votes appeared first. When sorting the threads by votes, students could choose to sort by number of 'up-votes,' which indicated student appreciation for the post, or 'down-votes,' which indicated their dislike for the post. There was also a search function enabling students to peruse the forum for topics of interest.

When we retrieved the data in November, a total of 96,696 posts were made by students during and subsequent to the course. For this study, we used only the 90,441 posts that were made during the 14 weeks of the course from March 4 through June 15, 2012. Of the 108,008 students who clicked at least once on the course site, 8,894, or slightly over eight percent of them posted at least one time on the discussion forum. The distribution of posts per student was positively skewed (Mean = 10.23, SD=44.66, Median = 2, Mode = 1), with the total posts per student ranging from 1 to 1442 for those who posted at all. Each student post was time-stamped and identified so that complete conversations could be re-constructed. The data were de-identified and the time stamp was removed prior to initiation of our analysis.

Phase 1 – Development of Coding Framework

Our research question, 'What is the topic of students' posts and the role assumed by posters on an open discussion forum?' provided the conceptual framework for this study. Our approach entailed classifying the topics into meaningful categories so that they could

be quickly located and clustered in ways that would facilitate future research. To accomplish this, we applied codes[1] that described the primary themes of the posts.

Our first task was to gain perspective regarding the appropriate number and types of categories that should be used to describe the posts. This inductive methodology, referred to as open coding, is commonly used at this stage to identify, describe, or categorize phenomena found in qualitative data (Corbin & Strauss, 1990). A committee of five researchers performed open coding on a series of randomly selected threads to gain a general awareness of the type of communication that occurred in this space. Our findings from this examination were wide-ranging; we noted student exchanges containing advice for studying course material, questions and responses related to understanding course content or accessing course materials, commiseration with fellow students about difficulties experienced in solving homework problems or in using the course technology, requests for group formation in other online sites, e.g., Facebook, and even comments related to students' physiological states, indicating hunger or sleep deprivation.

Following this first review, we developed a tentative list of descriptive codes that would encompass the topics and roles we had noted. Creation of codes that possess conceptual and structural order is important to the viability of a coding schema. Miles and Huberman (1994) emphasize that codes should relate to one another in coherent, study-important ways. We chose codes that we posited to be associated with our student outcomes of interest—persistence and achievement. We began with four codes for topic—content, user interface/course structure, community-building/interpersonal, and nonspecific affective—and two codes for role—help-seeker and help-giver.

After defining or operationalizing our codes, we trialed them on randomly selected posts. It immediately became apparent that we needed to expand this framework and refine our definitions. We added the code 'tangential topic' to describe posts in which students appeared to be transferring their understanding of circuits and electronics to other contexts. We also combined the two codes--community-building/interpersonal and nonspecific affective--to create one code--social/affective--because we found these posts to be closely related. We developed separate codes for user interface/course structure, which we renamed course website / technology and course structure/policies because there appeared to be two distinct types of posts and we hypothesized that each type may show different association with our outcomes of interest. We developed the codes 'missing data,' for posts had been removed by course staff, and 'non-English,' for posts that appeared in other languages. For role of the poster, we added 'other' as a code to be applied to posts in which students did not ask for or receive help or information, but instead expressed an opinion.

We elaborated upon our former and newly added definitions, and then trialed the revised framework on an additional 500 posts. This trial was conducted so that each of the five

[1] Codes are tags or labels that assign meaning to data collected during a study. They can be attached to chunks or pieces of information as short as one word or as lengthy as a paragraph. They can be simple and descriptive, such as a category label, or they can be more complex and inferential, such as ascribing a motive for behavior (Miles & Huberman, 1994).

researchers coded a total of 200 posts, with 100 of those overlapping posts coded by another researcher, and the remaining 100 overlapping those coded by a different researcher. This type of procedure, known as check-coding, allows researchers to identify code definitions that need elaboration, or even codes that need to be re-defined (Miles & Huberman, 1994). It also allowed us to evaluate inter-coder reliability or degree of agreement between coders. For this round of coding, the inter-coder reliability, calculated via percent agreement, ranged from 75% to 91% between the five possible combinations of researchers for topic of the post, and from 76% to 86% for role of the poster.

Following this trial, we each met with the two other researchers who coded the same posts to discuss discrepancies in our coding. We continued to refine our code definitions until we reached agreement on a framework that contained well-defined codes for all of the types of posts we encountered. During this process, we changed the 'tangential topics' code to 'other coursework' to categorize posts in which students made references to courses other than 6.002x. We found that identifying posts that discussed topics tangential to course content was difficult; we reasoned that identifying posts in which students referred to other courses may also allow us to determine some transfer of course understanding to other contexts. To further clarify the meaning of our codes, we found specific examples of posts related to each one and incorporated all of the information into a printed document that we referred to as our 'codebook.'

The product of our efforts was a two-dimensional coding framework that would classify posts according to topic of the post and role that the poster assumed. We defined eight subcategories for topic and three subcategories for role of poster as shown in Tables 1 and 2 (See Appendix A for further detail).

Table 1.

Topic of Post

Code	Definition
1. Content	Posts specifically addressing circuits and electronics material
2. Other coursework	Posts discussing courses other than circuits and electronics
3. Social/affective	Posts addressing social, emotional, or community-building aspects of the class
4. Course website/technology	Posts that addressed the online interface
5. Course structure/policies	Posts regarding the course organization, guidelines, or requirements
6. Other	Posts conveying anything not related to class content, other courses, social aspects of the class, course website or technology, or course requirements
7. Missing data	Posts in which data had been censored by course staff (in these cases, particular thread numbers could not be located)
8. Non-English	Posts written in other languages

Table 2.

Role of Poster

Code	Definition
1. Help-seeker (or Information-seeker)	Posts in which the poster asked for help, information, pointers, etc.
2. Help-giver (or Information-giver)	Posts in which the poster gave help, insight, or provided information
3. Other	Posts in which the poster was not explicitly seeking, declaring, or providing information, such as an opinion

As we trialed each iteration of the coding framework, we had to make several important decisions about our coding protocol. First, we found it necessary to allow dual codes to encompass the complexities of the topic and role dimensions. This meant that two codes could be applied to categorize the topic of one post, and two additional codes could be applied to describe multiple roles assumed by the poster. In a single post, it was not uncommon to find a response to another student's question about course content, an expression of emotion about understanding or not understanding that content, as well as a general question to other students regarding whether they would give a different answer to the question. In that illustration, the student's response to another student's question about content would be coded as 'help-giver' for role and 'content' for topic; their expression of emotion would be coded as 'social/affective' as a second topic; and their question about alternate answers would be coded as 'help-seeker' as a second role.

Our second important decision was that we would code posts independently of the thread in which they appeared. When students began a conversation about a particular topic, they started a new 'thread' on the forum. Each thread contained a series of posts, in which other students commented on or responded to the original question or statement. Our preliminary work involved coding only the first post of each thread because we believed this would identify the topic of that particular conversation. However, we found that one thread often contained multiple topics and that the topics were not always accurately identified in the first post. Often students' original statement of the problem was a misdiagnosis. The most common examples of this were threads in which students attributed difficulties with homework to technology errors that were actually their own errors in calculation or in understanding course content. In order to gain a more accurate sense of the number of posts in each of our coding categories, we elected to code at the individual post level rather than at the thread level. This decision also allowed us to gain more information regarding help- or information-givers because the first posts in threads were typically initiated by students asking questions, and the help- or information-giving role became more evident in subsequent posts within the same thread.

Phase 2 – Coding of Posts by Independent Coders

A further check of our coding framework was conducted by handing it off to independent coders to determine if our categories were comprehensive and our category descriptions provided adequate guidance for coding to those who were unfamiliar with the data.

Inherent in this phase was selecting the sample of posts to be used for coding. Realizing that it would be impossible for a human to code each of the 96,696 posts in any timely fashion, we chose to code slightly under five percent, or 4,500 posts. In our preliminary work, we noted that the central focus of students' posts changed as the course progressed. To capture this within the data, we divided the course into four time periods (quarters) and randomly selected 1125 posts from each quarter. We did not select posts that occurred after the end of the course, even though students continued to communicate with each other on the site well over a year after the course officially ended. We also noted that students tended to post more frequently at the beginning and end of the course than in the middle. To gain a more accurate estimate of posts described by each of our coding categories, we will weight the number of codes per category in each time period by the proportion of the total posts that occurred within that time period.

Two graduate students who were not on our original committee were recruited to code the 4500 posts. After their initial training and check-coding, these individuals would conduct the final coding of the data. The training consisted of orienting them to the codebook and discussing examples of each code. Following the training session, each coder was given 2500 posts, with 500 of the posts overlapping so that we could check for inter-coder reliability. The coders were instructed to code 32 of the overlapping posts from each of the four time periods (128 total for each coder), using codes from what we considered to be our final coding framework. Following completion of this first round of coding, we checked for inter-coder reliability by calculating percent agreement and found it to be only 59% for topic and role combined. We met with the coders to discuss difficulties in utilizing the codebook and to resolve disagreements. We made clarifications to the code definitions, and the coders then applied codes to another 128 posts. For this second round, they reached 81% agreement for topic, and 73% agreement for role of the poster. The coders then returned to the first 128 posts and re-coded, this time reaching 73% agreement for topic and 72% for role of the poster. Although there is no agreed upon threshold for appropriate inter-coder reliability (Campbell, Quincy, Osserman, & Pedersen, 2013), these percentages fall within a range considered acceptable for this type of data (Fahy, Crawford, Ally, Cookson, & Keller, 2000). Following the third round of coding and check-coding, the coders were instructed to code the remainder of their 2500 posts.

Lessons Learned

Utilization of the coding framework presented some particular challenges when implemented with the 6.002x discussion forum data. One significant challenge--that of interpretation based on the researcher's own bias--is inherent in all qualitative data analysis (Miles & Huberman, 1994) and was also present as we resolved differences between coders after they coded the same posts. In our research context, where posts were dissociated from the thread in which they occurred, it became very easy to infer the motive of the poster based on personal interpretations. For example, the role of the poster in the comment, 'I got 1.25 and it still isn't right,' could be viewed as information-giver, but it could also be viewed as an information- or help-seeker because the student obviously hadn't arrived at the correct answer. It was very tempting to speculate the circumstances under which this post was made. Yet another perspective might be to see this post as

social/affective, voicing frustration or confusion over repeated attempts to correctly solve the problem. We soon realized that from the latter perspective, many posts could be classified as social/affective. The fact that students were posting at all could be interpreted to mean that they were attempting to make a connection with other students.

Our resolution to these dilemmas was to clarify what our codes could or could not distinguish and what this meant for our future research questions. Some posts were simply difficult to put in a particular 'bin.' During our check-coding discussions, we began to approach these situations by thinking about what type of information would be needed to answer our future research questions and what was not relevant to code for. For example, when considering our future research question regarding factors that may be related to students' persistence or achievement in the course, we would not be able to accurately assess the relationship of social/affective communication to those outcomes if we had liberally assigned that code to any post in which we inferred that students were reaching out to make social connections. Similarly, attributing emotion or an affective component to a post due to our own interpretation of the poster's tone would also add error to our future analysis. We agreed to rely only on information overtly displayed in the post itself, which for the above examples, meant that the poster stated an emotion-laden phrase such as "I really love this course," or made an overt social gesture such as "Would anyone like to join a group that meets in person?" rather than to guess or infer anything about the context or motivation behind the post. Although we may have lost some information by making this choice, we felt that it strengthened the reliability of our coding in addition to reducing potential error from simply being wrong about what the poster intended.

Additionally, our experience reinforced the importance of orienting coders to the coding framework with actual data, not just for training purposes, but as a chance to get to know the character of the posts and understand how the abstract notions on which the codes are based are realized in the forum (i.e., how people ask for help, norms regarding off-topic postings, etc.).

Conclusion

The purpose of this paper was to describe the methodology used to confront the challenges of analyzing discussion forum data from the first edX MOOC, "Circuits and Electronics." The open, relatively unstructured nature of the 6.002x discussion forum resulted in student interactions that were extremely diverse with respect to content as well as nature of the interaction. Our efforts to answer several overarching research questions about the benefits of the discussion forum participation toward positive student outcomes began by answering the more specific question of what types of information students' actually posted in this space and what role or roles they assumed in these exchanges. We reviewed current recommendations for research into computer-supported collaborative learning and ascertained that our research questions addressed current needs in this area. We then developed a framework to classify copious amounts of data into a manageable number of categories so that further analysis could be conducted in targeted areas of interest. We detailed our methodology for development and testing of this coding framework and discussed challenges that arose during its implementation.

The results of our efforts enable us to provide an empirical account of how students actually use the discussion forum in a MOOC--information that may assist developers as they strive to enhance positive outcomes from students' use of this medium. Additionally, our work provides a structure that facilitates easy identification of data integral to our future research. Our next steps will be to explore characteristics of productive dialogue between students in this space, using a more in-depth coding schema to analyze students' posts. For this work, we intend to analyze entire threads, which will allow examination of a complete 'conversation' between students and will help to alleviate our earlier difficulties in coding ambiguous posts. We will also explore the relationships between the subject matter of students' posts, their role within those posts, and their achievement or persistence in the course.

We believe our framework is generalizable to other MOOC discussion forums, although modifications to the current framework may be desirable dependent upon the results of our future analysis. For example, if we find that the code 'references to courses other than 6.002x' has little association with any student outcome of interest, this code could be eliminated to reduce complexity of the coding process. As we probe more deeply into specific categories of posts, we may find it advantageous to develop subcategories for codes such as 'social / affective,' as the posts that contain social overtures may be associated with different student outcomes than those expressing positive or negative emotion. We hope that providing this level of transparency to our process of development will enable others to benefit from our experiences and build on our work. Additionally, this framework can provide the foundation for the future utilization of natural language processing to code discussion forum data.

Acknowledgements

This material is based upon work supported by the National Science Foundation Division of Research on Learning in Formal and Informal Settings under Grant #1258448. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The authors are also grateful for the support of the MIT RELATE group, edX, and the 6.002x course team. The authors would also like to acknowledge the contributions of Yoav Bergner and Samuel Finegold, who assisted with development of the coding framework, and those of Deepa Vasudevan, who assisted with testing the framework.

References

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*(3), 307–359.

Breslow, L., Pritchard, D.E., DeBoer, J., Stump, G.S., Ho, A.D., Seaton, D.T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research and Practice in Assessment, 8*, 13-25.

Campbell, J.L., Quincy, C., Osserman, J., & Pedersen, O.K. (2013) Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*(3), 294-320.

Corbin, J., & Strauss, A. (1990). Grounded theory method: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13,* 3-21.

DeBoer, J., Stump, G.S., Seaton, D., & Breslow, L. (2013, June). Diversity in MOOC students' backgrounds and behaviors in relationship to performance in 6.002x. *Proceedings of the Sixth Learning International Networks Consortium Conference,* Cambridge, MA.

De Smet, M., Van Keer, H., & Valcke, M. (2008) Blending asynchronous discussion groups and peer tutoring in higher education: An exploratory study of online peer tutoring behavior. *Computers & Education, 50*(1), 207-223.

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education, 46,* 6-28.

Fahy, P.J., Crawford, G., Ally, M., Cookson, P., Keller, V., & Prosser, F. (2000). The development and testing of a tool for analysis of computer mediated conferencing transcripts. *The Alberta Journal of Educational Research, 66*(1), 85-88.

Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013) Toward a script theory of guidance in computer-supported collaborative learning. *Educational Psychologist, 48*(1), 56-66.

Fransen, J., Weinberger, A., & Kirschner, P.A. (2013). Team effectiveness and team development in CSCL. *Educational Psychologist, 48*(1), 9-24.

Hull, D.M. & Saxon, T.F. (2009) Negotiation of meaning and co-construction of knowledge: An experimental analysis of asynchronous online instruction. *Computers & Education, 52,* 624-639.

Husman, J., Lynch, C., Hilpert, J., & Duggan, M. (2007, June). Validating measures of future time perspective for engineering students: Steps toward improving engineering education. Paper presented at the American Society for Engineering Education Annual Conference and Exposition, Honolulu, HI.

Järvelä, S. & Hadwin, A.F. (2013). New frontiers: Regulating learning in CSCL. *Educational Psychologist, 48*(1), 25-39.

Järvelä, S., & Häkkinen, P. (2002). Web-based cases in teaching and learning: The quality of discussions and a stage of perspective taking in asynchronous communication. *Interactive Learning Environments, 10,* 1–22.

Jeong, H., & Chi, M.T.H. (1997). *Construction of shared knowledge during collaborative learning.* Proceedings of the 2nd international conference on Computer support for collaborative learning, Toronto, Ontario.

Kortemeyer, G. (2006). An analysis of asynchronous online homework discussions in introductory physics courses. *American Journal of Physics, 74*(6), 526-536.

Kirschner, P.A., & Erkens, G. (2013). Toward a framework for CSCL research. *Educational Psychologist, 48*(1), 1-8.

Kizilcec, R.F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. LAK'13 Leuven, Belgium.

Layton, R.A., Loughry, M.L., Ohland, M.W., & Ricco, G.D. (2010). Design and validation of a web-based system for assigning members to teams using instructor-specified criteria. *Advance in Engineering Education,* 2(1), 1-28.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis.* Thousand Oaks, CA: Sage Publications.

Ohland, M.W., Loughry, M.L., Woehr, D.J., Finelli, C.J., Bullard, L.G., Felder, R.M., Layton, R.A., Pomeranz, H.R., & Schmucker, D.G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self and peer evaluation. *Academy of Management Learning & Education, 11* (4), 609-630.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (1999). Assessing social presence in asynchronous text-based computer conferencing. *Journal of Distance Education, 14*, 51–70.

Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research, 13*(1), 89–99.

Salovaara, H. (2005). An exploration of students' strategy use in inquiry-based computer-supported collaborative learning. *Journal of Computer Assisted Learning, 21,* 39-52.

Schellens, T. & Valcke, M. (2006). Fostering knowledge construction in university students through asynchronous discussion groups. *Computers & Education, 46*(4), 349 – 370.

Schrire, S. (2006). Knowledge-building in asynchronous discussion groups: going beyond quantitative analysis. *Computers & Education, 46*(1), 49–70.

Shell, D. F., Husman, J., Turner, J. E., Cliffel, D. M., Nath, I., & Sweany, N. (2005). The impact of computer supported collaborative learning communities on high school students' knowledge building, strategic learning, and perceptions of the classroom. *Journal of Educational Computing Research, 33*(3), 327–349.

Slof, B., Erkens, G., Kirschner, P.A., Jaspers, J.G.M., & Janssen, J. (2010). Guiding students' online complex learning-task behavior through representational scripting. *Computers in Human Behavior, 26,* 927-939.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science, 323*(5910), 122–124.

Song, L., & McNary, S.W. (2011). Understanding students' online interaction: Analysis of discussion board postings. *Journal of Online Learning, 10*(1), 1-14.

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research, 69*(1), 21–51.

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer supported collaborative learning. *Computers & Education, 46*(1), 71–95.

Weinberger, A., Ertl, B., Fischer, F., & Mandl, H. (2005). Epistemic and social scripts in computer-supported collaborative learning. *Instructional Science, 33,* 1-30.

Appendix A

Codebook for Discussion Forum Posts – 6.002x

(Final Version 7/10/2013)

INSTRUCTIONS

Use the codes below to code the 6.002x discussion forum data according to the following guidelines:

- Give each of the posts one code for 'topic' and one for 'role of poster.'

  *Exception - codes **3, 6, 7, and 8** require no coding for 'role.'

- You may give dual (2) codes for topic and two codes for role if appropriate. See examples of this on page 6.
- Use the 'Other' category **only** when all other possibilities have been exhausted.

The coding categories are as follows:

TOPIC

1. Content
2. Reference to courses other than 6.002x
3. Social/Affective
4. Course Website/Technology
5. Course Structure/Policies
6. Other
7. Missing data
8. Non-English

ROLE OF POSTER

1. Help-seeker or information-seeker
2. Help-giver or information-giver
3. Other

General Definitions: Topic of Post

1 Content: This category encompasses posts that address the subject matter in "Circuits and Electronics (6.002x)." This includes, but is not limited to, questions about the homework, questions to clarify understanding of concepts, additional insights on class topics, and comments regarding metacognition or self-regulation.

2 Reference to courses other than 6.002x: This category encompasses posts that specifically address other courses, whether or not they are related to "Circuits and Electronic (6.002x)."

3 Social/affective: This category encompasses posts that address the social or community building aspects of the class, enacting a social norm, or communicating an emotion. This includes, but is not limited to, posts in which students are looking to connect with "like" or "unlike" students, seeking connections with teachers or course staff, forming study groups, building communities, ritual communication, or expressing emotions such as gratitude, joy, anger, frustration, or confusion.

4 Course website/technology: This category encompasses posts that address the online interface, including but not limited to technical aspects of webpage accessibility and function.

5 Course structure/policies: This category encompasses posts that address the course organization and course guidelines, including, but not limited to sequence of topics, grading, deadlines, and earning certificates.

6 Other: This category includes anything that is not related to the class content, the social aspect of the class, course website or technology, or course requirements.

7 Missing data: Question/comment/answer has been removed from the discussion forum.

8 Non-English: Post is in a language other than English.

General Definitions: Role of Poster

***Note – topic codes 3, 6, 7, 8, require no codes for this category**

1 Help-seeker (or information-seeker): This category includes posts in which the poster is asking for some kind of help, information, pointers, etc.

2 Help-giver (or information-giver): This category includes posts in which the poster is giving help, insight, or providing information.

3 Other: This includes any posts in which the poster is not explicitly seeking, declaring or providing information, e.g., an expression or opinion.

Examples - Topic

1 Content: This category encompasses posts that address the topics in "Circuits and Electronics (6.002x)." This includes, but is not limited to, questions about the homework, questions to clarify understanding of concepts, additional insights on class topics, and comments regarding metacognition or self-regulation.

- "I think there is some error in question s1e9 that's why 2, 3, 4 answers are not matching, isn't it?"
- "What are high and low level noise margins?"
- "I will clarify how real-life circuit boards work."
- "Several people (including myself) have initially gotten the problem wrong, because we failed to notice that the expected answer is in milliwatts, not watts, so be careful!"
- "Are you forgetting to include the ground for 0 volt reference?"
- "Never mind, silly mistake on my part. The rules of mathematics still hold universally :)!"
- "One thing I'm doing while working on this course is to observe my own learning. Recalling my days in school, I seemed to remember learning more from books than from lectures. It does seem now that while I have learned a lot from our excellent video lectures, I still had to rely on the text to finally 'nail down' some of the material. It just seems more 'continuous' to me. It just seems to be the way that works best for me. And of course combining sources is very nice. I would be interested in hearing others' observations about how their favorite mode of learning feels to them. Why it works best for them. Why they favor videos (or not); why they favor discussions (or not), reading a book (or not), working problems (or not), and so on. Thanks." [This is an example of a post that can be given a second code for topic - "3"-social/affective.]

2 Reference to courses other than 6.002x: This category encompasses posts that specifically address other courses, whether or not they are related to "Circuits and Electronic (6.002x)."

- "Failure analysis and responsible failure reporting seems to be trending up lately. Does [name removed] incorporate failure studies into its later engineering courses?"
- "Are they offering 6.00x in the spring?"
- ""Many professions require continuing education, the time is measured in CEUs which are supposedly one contact hour each. I'm not sure if that's what the request is for but they are not the same as credit hours"
- "It would be a pleasure to have you on board. :) I hope to see an E&M course as well. Fingers crossed for the Fall. "[This can be dual coded for topic - "3"-social/affective.]

3 Social/affective: This category encompasses posts that address the social or community building aspects of the class, enacting a social norm, or communicating an emotion. This includes, but is not limited to, posts in which students are looking to connect with "like" or "unlike" students, seeking connections with teachers or course staff, forming study groups, building communities, ritual communication, or expressing emotions such as gratitude, joy, anger, frustration, or confusion.

- "I love circuits!"
- "MOSFETS are hard!"
- "I'm new here, please help"
- "Is there anyone that could possibly explain the lecture to me? ...I have no idea about this stuff. If possible could someone be sort of like my tutor. I know none of you know me, but I would appreciate the help."
- "Please share here only educational email id which use only for education. Thanx. mine is [name deleted]@gmail.com"
- "I think I got a certificate. I won't say what grade, but you don't expect to become a concert pianist in 14 weeks, even if a great instructor has shown you exactly what you have to do. Now to go back over everything, do some of the problems at the ends of the chapters, list the Aha moments, redo homework and exams, but at a pace which allows me to smell the roses. .. … And in addition I'll be able to tell [university name deleted] how to structure their online Russian Literature Course. 6.002x has opened up a whole new future!"
- "@[name deleted], quit apologizing, it doesn't suit you. You have no reason to apologize.
- "Your English is excellent and so are your answers."
- "It looks good. Thanks."

4 Course website/ technology: This category encompasses posts that address the online interface, including but not limited to technical aspects of webpage accessibility and function.

- "I see the page with the YouTube links, and got jdownloader but can someone tell me how to actually use the YouTube link page to make Jdownloader get all the videos?"
- "If I correct error in homework (red x to green tick) immediately - is it considered ok for grading? You know all this number crunching can get flaky at times."
- "[language deleted]characters not showing up properly in certificate."
- "You will probably remember that Backspace will remove a circuit element and that R will rotate it. I have discovered that Control-C will copy, Control-V will paste and Control-X will remove selected elements…Do you have any other useful comments about Sandbox?"
- "MIX staff - Another day (4/16/2012) with lots of recurring 'Internal Server Errors.' I've gotten two so far, one on a forum answer, the other on a forum comment. Other users

have experienced the same thing today. Sure hope this doesn't happen during the exam!"
- "People have been asking this a lot. It seems that they've been having some trouble generating the certificates. Check out this link for more info: [web link deleted]. And of course, just type 'certificates' on the search bar if you want to catch up with the discussion. Cheers."
- "How do I see what was my grade?? Help please!!"

5 Course structure / policies: This category encompasses posts that address the course organization and course guidelines, including, but not limited to sequence of topics, grading, deadlines, and earning certificates.

- "Hello I only joined on the 19th of March, however there was a homework due on the 18th what do I do for this? Should I still fill it in or just leave it; will it even get a mark?"
- "I submitted an answer to Q3 part 2 of the Midterm after my 24 hours had elapsed but it graded the question and seems to have increased my grade in the course. I was expecting it to just let me see if my new answer was correct, not change my grade. Can you please reduce my grade back to 96%, which was what it was at when 24 hours had elapsed?" [This can be dual coded for topic - "4"-course technology/website]
- "For now we give a 24 hour window for submissions. So stick to the due dates at your local time and you should be fine."
- "Just wondering considering many questions are being asked related to HW, but nothing related to Tutorials.... so again Does anyone care about Tutorials??"
- "Guys I finished the final exam and I knew that MIT will need 24 hours to give certificates...Where is mine??"

6 Other: This category includes anything that is not related to the class content, the social aspect of the class, course website or technology, or course requirements.

- "If you're a [social network] user, do yourself a favor and change your password right now — according to a new report from Dagens IT, nearly 6.5 million encrypted [social network] passwords were recently dumped onto a hacker forum."
- "A correction: Coursera is not a consortium. They're a VC-funded startup, like Udacity. They did not reveal this to students until recently, but they are and have been a for-profit."

7.0 Missing data: Question/comment/answer has been removed.

8.0 Non-English: Post is in a language other than English.

Examples – Role of Poster

1 Help-seeker (or information-seeker): This category includes posts in which the poster is asking for some kind of help, information, pointers, etc.

- "What are high and low level noise margins?"
- "I see the page with the YouTube links, and got jdownloader but can someone tell me how to actually use the YouTube link page to make Jdownloader get all the videos?"
- "In other words, I seem to have the correct values, but can't reproduce the plots shown..."

2 Help-giver (or information-giver): This category includes posts in which the poster is giving help, insight, or providing information.

- "I will clarify how real-life circuit boards work."
- "Several people (including myself) have initially gotten the problem wrong, because we failed to notice that the expected answer is in milliwatts, not watts, so be careful!"
- "Are you forgetting to include the ground for 0 volt reference?"
- "And I was talking about the graph there... ;o)"
- "Not only is it more linear placed before the amplifier, but you avoid the need for inductors (which are the worst behaving passive circuit element). And you can let the low output impedance of the power amplifier dominate the speaker coil directly and prevent resonances between the speaker system and the filter components."
- [web link removed]

3 Other: This includes any posts in which the poster is not explicitly seeking, declaring or providing information, e.g., an expression or opinion (and if topic = "1," "2," "4," or "5").

- "One thing that amazed me of this course is the embedded Circuit Sandbox. First time seeing one on a browser that can do DC, AC and even transient analysis. And there are all basics components further than just resistor, capacitor and inductor."
- "I would caution you that as a first year undergraduate student, you are unlikely to have taken differential equations. Based on the syllabus, I would consider a first course in differential equations to be either a pre-requisite, or a co-requisite for successful completion of the entire 6.002x program."

Examples of Dual Codes:

- "One thing I'm doing while working on this course is to observe my own learning. Recalling my days in school, I seemed to remember learning more from books than from lectures. It does seem now that while I have learned a lot from our excellent video lectures, I still had to rely on the text to finally 'nail down' some of the material. It just seems more 'continuous' to me. It just seems to be the way that works best for me. And of course combining sources is very nice. I would be interested in hearing others' observations about how their favorite mode of learning feels to them. Why it works best

for them. Why they favor videos (or not); why they favor discussions (or not), reading a book (or not), working problems (or not), and so on. Thanks."
[This is Topic = "1"Content and "3" Social/affective (enacting a social norm); Role = "1" Information-giver and "2" Information seeker.]

- "Sqrt[3] is a solution to the first equation. The fact that Mathematica didn't come up with that is what prompted my question. Evaluate [Abs[1/(1 + j*3^(1/2))]] outputs 1/2. But thanks for the information regarding Abs of complex numbers."
  [This is Topic = "1" Content, and "3" Social/affective; Role = "1" Help-giver.]
- "Could you identify which exercise, homework, or lab problem you're working on?"
  [This is topic = "1" -content; Role = "1" Information-seeker, and "2" Help-giver. Dual codes are given for role in this case because it is unknown if the poster also has a question about the content or if he/she is attempting to focus the question in order to provide assistance.]