International Review of Research in Open and Distributed Learning

Cross Validating a Rubric for Automatic Classification of Cognitive Presence in MOOC Discussions

Yuanyuan Hu, Claire Donald et Nasser Giacaman

Citer cet article

Hu, Y., Donald, C. & Giacaman, N. (2022). Cross Validating a Rubric for Automatic Classification of Cognitive Presence in MOOC Discussions. *International Review of Research in Open and Distributed Learning*, *23*(2), 242–260. https://doi.org/10.19173/irrodl.v23i3.5994

Résumé de l'article

As large-scale, sophisticated open and distance learning environments expand in higher education globally, so does the need to support learning at scale in real time. Valid, reliable rubrics of critical discourse are an essential foundation for developing artificial intelligence tools that automatically analyse learning in educator-student dialogue. This article reports on a validation study where discussion transcripts from a target massive open online course (MOOC) were categorised into phases of cognitive presence to cross validate the use of an adapted rubric with a larger dataset and with more coders involved. Our results indicate that the adapted rubric remains stable for categorising the target MOOC discussion transcripts to some extent. However, the proportion of disagreements between the coders increased compared to the previous experimental study with fewer data and coders. The informal writing styles in MOOC discussions, which are not as prevalent in for-credit courses, caused ambiguities for the coders. We also found most of the disagreements appeared at adjacent phases of cognitive presence, especially in the middle phases. The results suggest additional phases may exist adjacent to current categories of cognitive presence when the educational context changes from traditional, smaller-scale courses to MOOCs. Other researchers can use these findings to build automatic analysis applications to support online teaching and learning for broader educational contexts in open and distance learning. We propose refinements to methods of cognitive presence and suggest adaptations to certain elements of the Community of Inquiry (CoI) framework when it is used in the context of MOOCs.

Cet article est diffusé et préservé par Érudit.

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche.

https://www.erudit.org/fr/

Cross Validating a Rubric for Automatic Classification of Cognitive Presence in MOOC Discussions

Yuanyuan Hu, Claire Donald, and Nasser Giacaman
Faculty of Engineering, The University of Auckland

Abstract

As large-scale, sophisticated open and distance learning environments expand in higher education globally, so does the need to support learning at scale in real time. Valid, reliable rubrics of critical discourse are an essential foundation for developing artificial intelligence tools that automatically analyse learning in educator-student dialogue. This article reports on a validation study where discussion transcripts from a target massive open online course (MOOC) were categorised into phases of cognitive presence to cross validate the use of an adapted rubric with a larger dataset and with more coders involved. Our results indicate that the adapted rubric remains stable for categorising the target MOOC discussion transcripts to some extent. However, the proportion of disagreements between the coders increased compared to the previous experimental study with fewer data and coders. The informal writing styles in MOOC discussions, which are not as prevalent in for-credit courses, caused ambiguities for the coders. We also found most of the disagreements appeared at adjacent phases of cognitive presence, especially in the middle phases. The results suggest additional phases may exist adjacent to current categories of cognitive presence when the educational context changes from traditional, smaller-scale courses to MOOCs. Other researchers can use these findings to build automatic analysis applications to support online teaching and learning for broader educational contexts in open and distance learning. We propose refinements to methods of cognitive presence and suggest adaptations to certain elements of the Community of Inquiry (CoI) framework when it is used in the context of MOOCs.

Keywords: cognitive presence, MOOC, text classification, online discussion, content analysis

Introduction

The Problem

In this paper, we argue that the existing empirical inquiries and theoretical frameworks for analysing learning engagement in conventional for-credit university courses are limited when analysing learning engagement in massive open online courses (MOOCs), due to the differences in their educational contexts. MOOCs differ from traditional, smaller-scale online courses in terms of the course design, with shorter course durations, limited direct educator involvement (Kovanović et al., 2018), a wide range of learner profiles, and diverse learner motivations (Alario-Hoyos et al., 2017). During the COVID-19 pandemic, MOOCs gained much more attention in addressing the limitations of remote learning, as they provide learners with a diverse range of educational experiences (Buchem et al., 2020; Cha & So, 2020). MOOC educators require support to monitor and moderate learner progress in these massive audiences, and MOOC learners require responsive, high-quality feedback and remediation from educators. Automatic classification of cognitive presence (to supplement automated feedback which is currently the norm) can provide such pedagogical support. A vital foundation for such a classification system is a reliable theoretical basis.

The asynchronous discussion forum in MOOCs offers a virtual zone for participants to interact mutually through written dialogue. These written conversations, or messages, provide educators and researchers with meaningful insights into learners' critical discourse (i.e., critical thinking, higher-order thinking, and cognitive presence). However, experimental studies are required to validate and improve the methods of critical discourse analysis in online discussion forums for broader educational contexts, such as MOOCs (Amemado & Manca, 2017; Kaul et al., 2018). Garrison et al. (1999, 2001) proposed the community of inquiry (CoI) framework and its coding rubrics to evaluate cognitive presence and two other dimensions in online transcripts by content and textual analysis methods. Over two decades, this framework has been broadly used to assess students' learning and guide learning designs in traditional, online, for-credit, smaller-scale courses (Liu & Yang, 2014; Sadaf & Olesova, 2017). There are still shortcomings for the CoI coding rubrics to reliably assess critical discourse in MOOCs. For example, clear instances of the cognitive presence phases have not been elucidated by Garrison et al. (2001); therefore, researchers have had to revise the rubric each time they used it (Rourke & Kanuka, 2009). Also, the coding rubric was initially developed as a descriptive, qualitative analysis method in smaller-scale courses rather than as a quantitative, inferential procedure (Garrison, 2007). Moreover, online discussion consists of an informal and conversational flow, which is relatively chaotic and does not fit into the coherent patterns in the CoI framework (Xin, 2012).

The Significance

The validation of a cognitive presence rubric in MOOCs is a crucial foundation for developing automatic approaches for real time learning support and remediation. Preparing reliable and valid machine learning data sets is an essential prerequisite for training automatic classifiers (Ullmann, 2019). Analysing the common language patterns in the data sets is beneficial for selecting appropriate machine learning algorithms and predictive features (Mladenić, 2010). Automatic evaluation of learners' cognitive presence in MOOCs can help educators monitor the learners' progress in real time and provide personalised feedback

at scale. From a learner's perspective, effective and efficient feedback from both educators and peers encourages high participation in MOOCs, assisting students to achieve their learning goals (Phan et al., 2016).

Purpose and the Research Questions

This study aims to cross validate the use of an adapted coding rubric (Hu et al., 2020) to categorise online discussion transcripts from a target MOOC into phases of cognitive presence. A larger dataset and more coders were involved in examining whether the inter-rater reliability could still reach excellent agreement, as reported in the previous study using less data and fewer coders (Hu et al., 2020). The disagreements between coders were also deeply analysed to gain insights about the feasibility of the cognitive presence phases in the CoI framework. Our main research question was: Is the adapted coding rubric a reliable tool to classify cognitive presence in MOOC discussions? The following sub questions were included to guide the main research question:

SQ1. What are the inter-rater reliability values when we classify the discussion messages from a target MOOC with more coders and a larger dataset than in the previous study (Hu et al., 2020)?

SQ2. What is the proportion of disagreements across all cognitive phases between coders, and what are possible causes of the disagreements?

# Review of Previous Studies

Cognitive Presence

The Community of Inquiry (CoI) framework proposed by Garrison et al. (1999) has been most widely used to analyse learning in online discussions for over two decades. It describes the educational experience that occurs in a learning community, in which "a group of individuals who collaboratively engage in purposeful critical discourse and reflection construct personal meaning and confirm mutual understanding" (Garrison & Anderson, 2011). The CoI framework has three dimensions, called presences: cognitive, social, and teaching presence. Cognitive presence, a primary dimension of the CoI, represents the critical reflection of knowledge (re)construction and problem-solving processes in the learning community (Garrison et al., 2001). Social presence manifests as social communication and emotional interaction between participants, which enriches learning outcomes. Teaching presence, the third dimension, describes the purposeful activities that direct and intervene with the learner's knowledge construction.

This study focused on analysing cognitive presence in MOOC discussion transcripts since the cognitive presence is the "primary issue" of students' learning evidence to be explored before other dimensions (Rourke & Kanuka, 2009). The other two presences of the CoI will be investigated in further research. We adopted the associated four phases of cognitive presence, (a) triggering event, (b) exploration, (c) integration, and (d) resolution from Garrison et al. (1999, 2001). The four phases (Figure 1), called the practical inquiry model, were borrowed from Dewey (1933) who originally described the steps of a complete thought. The definition of the four cognitive phases corresponding to the analysis of discussions in our

target MOOC is explained below.

Figure 1

The Practical Inquiry Model Showing the Four Phases of Cognitive Presence in a Learning Community
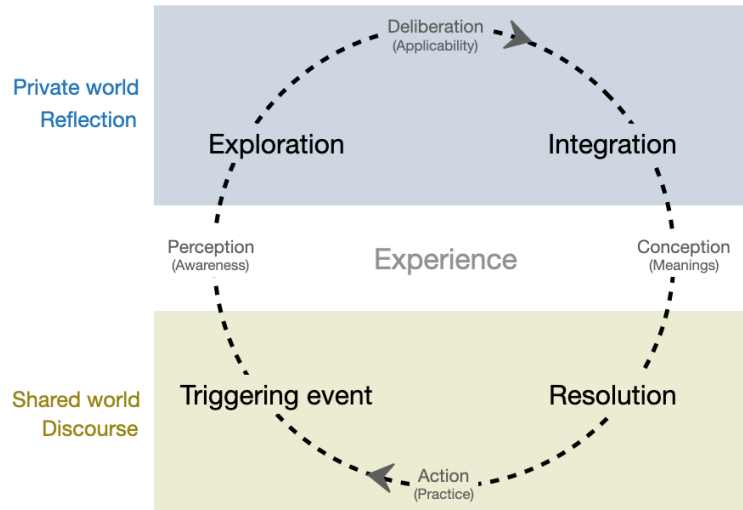
Note. Adapted from "Critical thinking, cognitive presence, and computer conferencing in distance education," by D. R. Garrison, T. Anderson, and W. Archer, 2001, *American Journal of Distance Education*, *15*, p. 9. Copyright 2001 by University of Chicago Press.

Analysis of Cognitive Presence Phases

Scholars have applied the classification rubric developed by Garrison et al. (2001) and the revised version (Park, 2009) to assess the quality of critical discourse in multidisciplinary online courses. More recent studies have used transcripts data from for-credit, online university courses, but few so far have used transcripts data to analyse cognitive phases in MOOCs. Table 1 summarises the prior research that we have reviewed.

Table 1

Summary of Prior Research Reviewed

Study	Year	Course context	Discipline	Messages *n*	Coders *n*	Agree-ment %	κ
Garrison et al.	2001	Graduate	Health	24	2	83.33	0.74
Kanuka et al.	2007	Undergraduate	Education	1014	2	-	0.57
Park	2009	Graduate	Nursing	Thematic segments	2	82.34 76.48[a]	-
Liu & Yang	2014	Undergraduate	Information ethics	200	3	90	-

Corich et al.	2006	Undergraduate	Computing systems	484 sentences	2	87	-
McKlin	2004	Undergraduate	History and political science	1300	6	74	0.60
Kovanović et al.	2014	Graduate	Software engineering	1747	2	98.1	0.974
Neto et al.	2018	Undergraduate	Biology	1500	2	91.4	0.86
Kaul et al.	2018	MOOC	Education	78	2	86	0.82
Hu et al.	2020	MOOC	Philosophy	1002	2	95.4	0.93

Note. [a] The Park study included two trials.

Garrison et al. (2001) first proposed their CoI model and reported their manual classification rubric with data from two online graduate-level courses. One course was on workplace learning and the other on health promotion. The former dataset (21 messages) was used to fine tune the measurement rubric, and the latter (24 messages) was used to report the inter-rater reliability result. They reached an agreement of 83.33% and Cohen's κ coefficient (Cohen, 1960) of 0.74 between two coders. It indicated that this classification rubric could be used to evaluate the quality of cognitive presence. However, the sample size was too small (i.e., under 100 messages) and would need further verification with larger, more diverse learner datasets to be sufficiently generalisable.

Two studies made their measurement unit at the message level following Garrison et al.'s (2001) method. Kanuka et al. (2007) used Garrison et al.'s (2001) rubric to analyse online discussion transcripts in an undergraduate education course. They achieved a very low Cohen's κ of 0.57 (1,014 messages) with two coders. The second study (Liu & Yang, 2014) used the rubric in an undergraduate information ethics course, with a percentage agreement of 90% (200 selected from 1,058 messages) reached by three coders. However, Park (2009) deemed that the classification of cognitive phases should be based on the unit of meaning or thematic level rather than the message level. Park (2009) also revised Garrison et al.'s (2001) rubric to use in an online graduate nursing course, reaching agreements of 82.34% and 76.48% in two trials with an unknown number of messages posted by 12 students.

Some studies analysed cognitive phases manually in preparation for developing automated classifiers. Corich et al. (2006) labelled cognitive phases in discussion transcripts from an online undergraduate computing systems course. The correlation between the two coders was 87% (104 messages of 484 sentences) using Garrison et al.'s (2001) rubric. This study was done at the sentence level instead of the message level. McKlin (2004) collected 1,300 learner messages from online undergraduate courses in history and political science. Researchers used 100 of 1,300 messages to report the measure of inter-rater reliability, and the remaining 1,200 messages were coded by six coders (200 messages each). The study reported 74% agreement and an average Cohen's κ of 0.60 (100 messages). Kovanović et al. (2014) classified 1,747 messages from an online graduate course in software engineering research, reaching a very high agreement of 98.1% and Cohen's κ of 0.974 between two coders. This dataset has been reused in four further studies to develop automated classifiers of cognitive presence. To analyse cognitive presence in transcripts written in other languages, Neto et al. (2018) used a Portuguese dataset from an undergraduate biology

course, with an agreement of 91.4% (1,500 messages) and Cohen's κ of 0.86 between two coders. This dataset was also used to evaluate the performance of transferring automated classifiers across languages (Barbosa et al., 2020).

Two studies with manual coding approaches of cognitive presence in the CoI focused on MOOCs. Kaul et al. (2018) applied Garrison et al.'s (2010) classification rubric to 78 messages from an education MOOC in the subcategories of all three CoI presences. Agreement between the two coders was 46% and 86% before and after coders' negotiations. An adapted coding rubric (Hu et al., 2020) of cognitive presence was proposed when Garrison et al.'s (2001) classification rubric could not encompass all the cognitive phases in online discussions from a philosophy MOOC. Two expert coders reached a percentage agreement of 95.4% (1,002 messages) and Cohen's κ of 0.93 using the adapted classification rubric. Both studies indicated that most of the coders' disagreements occurred between the exploration and integration phases when classifying cognitive presence phases.

Differences Between the Prior Work and This Study

The cognitive presence analysed in most of the studies we viewed was in the context of smaller-scale, for-credit university courses. Garrison et al.'s (2001) classification rubric preceded MOOCs by seven years. The first MOOC was developed in 2008 (Siemens, 2013). The wide range of learner demographics and diverse learner motivations cause the differences between MOOCs and traditional university courses. The typical MOOC audiences are mature adult learners who are employed and have tertiary qualifications (Dillahunt et al., 2014). Their motivations for learning are updating knowledge, personal curiosity, and upskilling themselves professionally (Alario-Hoyos et al., 2017). These differences may also impact the language they use in online discussions. For instance, students tend to write formally when participating in discussion forums in smaller-scale courses and in professional development MOOCs that are credit-bearing. In contrast, many MOOC learners tend to use a more conversational style of writing when they engage with MOOCs for less formal purposes than professional development or accreditation. We wondered whether the differences in educational contexts would impact the analysis of cognitive presence in discission transcripts.

The reliability reported in most of the reviewed studies was based on two coders. Although six coders were employed in McKlin's (2004) study, they only labelled 100 same messages. Similarly, in Neto et al.'s (2018) study, a third coder was only responsible for resolving disagreements (129 messages) between the other two coders. Although the previous study (Hu et al., 2020) reported excellent inter-rater reliability between two coders, the agreement outcome was reported after the coders' negotiations. We wondered whether the coding rubric of cognitive presence could be reliably applied to analyse MOOC discussions, when we enlarged the dataset, invited more coders to become involved, and reported the outcomes before coders' negotiations.

Methods

Data Description

The MOOC discussion data used in our study was from an archived run of the Logical and Critical Thinking (LCT) MOOC on the FutureLearn platform (University of Auckland, n.d.). It was an introductory undergraduate philosophy course designed and taught by a course design team at our university. This course taught basic concepts in logical and critical thinking (e.g., premises, arguments, etc.), linking those concepts with life experiences. The average number of enrolled users was approximately 11,000, and the discussion transcripts (comprising posts and their replies) included approximately 12,000 messages per course run. There were eight weekly topics with learning tasks in each course run. Firstly, sixteen tasks (two for each week) were evenly and randomly selected. Then, a sample of approximately 100 to 200 messages was randomly selected from each of the 16 tasks. We kept the entire sequential structure of each selected conversation instead of segmenting them to achieve an exact number. Totally, 1,917 messages were selected for this study.

The three coders were postgraduate students from the Philosophy Department, who were also the teaching assistants for the LCT MOOC. They were trained round by round (50 new messages for each round) before reaching agreements over 80% independently without negotiations. They reached an 81% agreement in the third round, so they were allocated to classify the 1,917 messages manually and independently based on the adapted rubric (Hu et al., 2020) The overall study proposal received ethical approval from the University Human Participants Ethics Committee.

Definition of Cognitive Presence Phases in This Research

The five categories of cognitive presence, including four processing phases and the "other" phase, are listed in Table 2. We provide a brief definition and a message example from the LCT MOOC for each category. These definitions are derived from Garrison et al. (2001), Hu et al. (2020) study, and learner messages in the LCT MOOC, and therefore influenced by the disciplinary context of this course, which is philosophy. More details about the definitions can be found in Hu et al. (2020) study.

Table 2

Definitions with Message Instances of the Five Categories in Cognitive Presence

Phase ID	Cognitive phase	Brief definition	Message example
1	Triggering event	Messages state users' confusion.	*"I do find it difficult to override over 30 years of the normalisation of poorly constructed sentences."*
2	Exploration	Messages provide information about the cause of the confusion but without a coherent conclusion.	*"Both overthinking and underthinking leads you to live in low levels of consciousness. I think that [one of the users] explains very well how to find the spot between the two approaches."*

3	Integration	Messages propose coherent conclusions to improve the situation with sufficient substantiation.	*"I think this counter argument works colloquially but not technically it doesn't follow from the premises that having a job will stop you wanting an iPhone unless you add an implied premise to that effect."*
4	Resolution	Messages apply, test, or argue the previous conclusions, usually as new constructs.	*"Another way to test it would be to see if similar positions eg heads of industry are also held by more left-handed people than statistics would suggest. It would be incredibly difficult to iron out other possible factors…"*
0	Other	Messages that do not fall into any category.	*"Thanks. I start that Mooc in May."*

Classification Process

After they were trained, the three coders used our rubric of cognitive phases to classify the sample data (1,917 messages) independently. The unit of analysis was on the message level since the classification on the theme and sentence level may have ignored the contextual information before and after the segment (i.e., theme or sentence) within an entire message. Multiple phases of cognitive presence sometimes existed in one message simultaneously, for example, when a learner stated, diagnosed, and resolved a question in a single post. Our coders labelled each message with the highest cognitive phase in that message, even when lower phases were also represented. An example message of this is:

> Initially, I too felt the conclusion might be that a revolt was required. However, the letter states that a revolt would be a way to get the council to listen. It's the same as cider vinegar would be a way to get rid of your dogs fleas. Therefore, it's a statement rather than an argument.

The first sentence indicates the learner's difficulty, which can be categorised as triggering event. The second sentence illustrates that the learner provided more information for diagnosing the difficulty, which can be the exploration phase. Then, the last two sentences made an analogy and drew a conclusion supported by the reasons stated previously, which can be labelled integration. In this case, the message was classified into the highest phase, integration, rather than the other two phases. More details about the coding rubric and message examples can be found in the adapted rubric of cognitive presence in MOOCs (Hu et al., 2020).

Results

The Overall Analysis of Cognitive Phases

The overall percentage agreement was 77.15%, where all three coders independently agreed on the labels for 1,479 of the 1,917 messages. The average Fleiss' κ (a statistical measure for categorical ratings between more than two coders) was 0.763, shown in Table 3. Across five categories of cognitive presence, the triggering event phase accounted for the highest agreement (κ = 0.828), followed by the phases "other" and exploration. There was less agreement among the coders on the higher cognitive phases of integration and

resolution. The average agreement between coder 1 and coder 2 reached Cohen's κ of 0.842. This result was higher than the other two combinations, where Cohen's κ was 0.704 between coder 2 and coder 3, and 0.744 between coder 1 and coder 3.

Table 3

Inter-Rater Reliability Among Three Coders Across Five Categories

Phase ID	Cognitive phases	Fleiss' κ
0	Other	0.792
1	Triggering event	0.828
2	Exploration	0.776
3	Integration	0.689
4	Resolution	0.667
Average		0.763

Note. N = 1,917 messages.

Table 4 illustrates the proportion of the five cognitive phases in the messages of agreement between coders. The phase of exploration accounted for most of the messages (55.46%), which far surpassed triggering event and integration. The highest phase, resolution (2.43%), and the lowest "other" phase (5.75%) had the smallest proportion of messages.

Table 4

Messages of Agreements Between Three Coders by Cognitive Phases

Phase ID	Cognitive phase	Messages	
		n	%
0	Other	85	5.75
1	Triggering event	279	18.86
2	Exploration	835	56.46
3	Integration	244	16.50
4	Resolution	36	2.43%

Note. N = 1,479 messages.

Disagreements Between Coders

In addition to agreements between coders, the distribution of disagreements (i.e., messages that were labelled differently by the coders) is worth considering. Table 5 describes the proportion of the disagreements between the three coders across different combinations of cognitive phases. Most of the disagreements had two labels rather than three (the latter was less than 1.5% in total). The proportion of inter-rater disagreements (96.13%) between adjacent cognitive phases far surpassed the non-adjacent

combinations (i.e., exploration and resolution, and "other" and exploration).

**Table 5**

Distribution of Disagreements

Combination of phases	Messages	
	n	%
Other (0) & triggering event (1)	66	15.07
Triggering event (1) & exploration (2)	78	17.81
Exploration (2) & integration (3)	227	51.83
Integration (3) & resolution (4)	50	11.42
Other (0) & exploration (2)	7	1.60
Exploration (2) & resolution (4)	4	0.91
Three labels	6	1.37

Note. N = 438 messages placed in multiple categories. Numbers in parentheses indicate the study's phase ID.

Another way to understand the distribution of the agreements and disagreements data is by using the five contingency tables in Figure 2, which align with Table 4 and Table 5. Figure 2 describes the distribution of messages classified by coder 1 and coder 2 when coder 3's labels ranged from the "other" phase to resolution, respectively. The blue cell on the diagonal of each table (e.g., 85 on the first table) represents the number of agreements between the three coders in each cognitive phase. The red cells (e.g., 38, 3, and 2 on the first table) demonstrate the number of disagreements.

Figure 2

Distribution of Five Cognitive Phases Between Three Coders
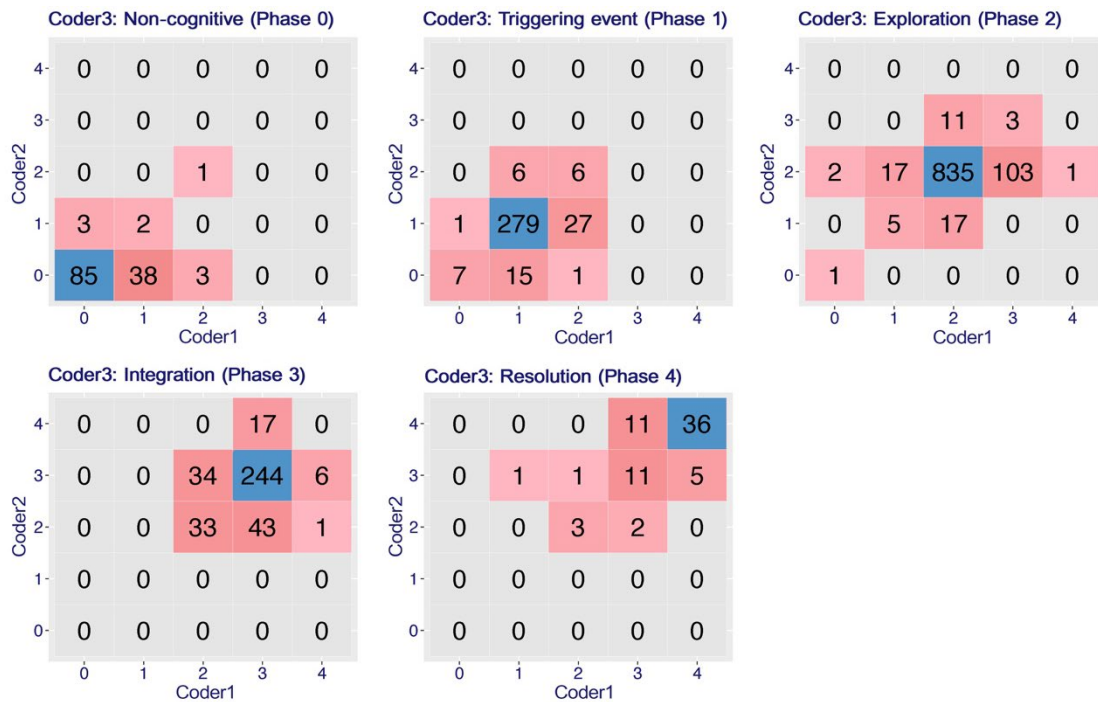
Note. Blue indicates the number of agreements among the 3 coders. Red indicates the number of disagreements. The red-colour scale is used to represent the disagreement cells. The larger the number, the darker the cell.

Discussion

Validation of the Adapted Coding Rubric—*SQ1*

The results (agreement of 77.29% and Fleiss' κ of 0.763) answer our first sub-question and show that a reliable inter-rater agreement was reached between the three coders in this research, as over 0.75 represents excellent agreement (Fleiss et al., 2003). The exploration phase accounts for most of the messages, and the resolution phase accounts for the least. This distribution rate is similar to the proportional results of cognitive phases from the reviewed studies (Kaul et al., 2018; Kovanović et al., 2014; Neto et al., 2018; Park, 2009). These discussion messages were reliably classified through the manual categorisation process, providing us with a clean training data set to develop automatic classifiers in our future work.

Disagreements Between Coders and What Caused the Disagreement—*SQ2*

Analysing the disagreements between coders (438 messages) can help to answer our second sub-question. Most of the three coders' disagreements appeared on adjacent phases of cognitive presence (Table 5 and Figure 2), in line with the findings in the previous study (Hu et al., 2020) with two coders. We analysed the common language patterns which may have caused disagreements between the coders.

Common Language Patterns of Messages Between Adjacent Phases

After reviewing the 66 messages (Table 5) which the coders labelled as "other" and triggering event, we came up with two possible reasons that may have caused these disagreements. First, messages with incomplete segments or concise sentences may have made it difficult to grasp the writer's purpose and meaning. There were ambiguous interpretations from different coders, even when they checked the previous and subsequent messages. The message instances were (a) "That works for me," and (b) "I am guessing…". Second, sentences using the structure "I like…" may have caused confusion. The verb *like* could be defined as either "I agree with you" or "appreciate". The instances were (a) "I like your worded comment. Nice!" and (b) "I like the questions you mentioned". The former could have been an indicator, "simple agreement," of a triggering event, whereas the latter could have been a predictor of social expression, which is part of the "other" phase.

The most common pattern reflected in the messages with both triggering event and exploration labels (78 messages as shown in Table 5) was the use of questions to deliver outside information or make personal claims. These language patterns confused our coders. "Ask questions" is a core indicator of the triggering event phase; however, learners can also propose ideas using sentences ending with question marks, such as in the case of rhetorical questions. These messages can be interpreted as an indicator of "suggestion for consideration" which is part of the phase of exploration. Two examples were: (a) "A good example of the strawman fallacy on me?" and (b) "At what point? Early on it is said 'will come to some philosophy department meeting'. Is that when?" Unlike writing essays or research reports as assignments in for-credit university courses, many online conversations in MOOCs contain informal writing. Coders cannot acquire sufficient information from the informal language patterns to verify the writer's actual intention. This vital difference in MOOC conversations compared to formal coursework may be associated with certain learner motivations, such as updating knowledge voluntarily and personal curiosity. This finding has a significant implication for building automatic classifiers of cognitive phases in future studies. For example, the number of question marks in a message in a MOOC would not be a positive indicator of the triggering event phase. However, in online discussions in smaller-scale, for-credit courses, the opposite can be true, and the number of question marks in a message may be reliably used as a predictor when building automatic classifiers (Kovanović et al., 2016).

The central debates in the 227 messages (Table 5) labelled both exploration and integration had two aspects. First, messages that contained conclusions with reasons raised a dispute about whether the supporting ideas were sufficient. A significant criterion to differentiate integration from exploration is that the message should reach "a coherent conclusion" by offering "sufficient substantiation" in the classification rubrics (Garrison et al., 2001; Park, 2009). This criterion is subjective and domain specific. Messages that provide solutions and implicit conclusions ending with a tentative phrase imply more of a "suggestion for consideration" (which should be labelled exploration), rather than sufficiently supported integration. For example, in our study, we saw this message: "I do not think it is an argument because a rates revolt is only a suggestion. The writer states that it is one way to make the councillors listen but does not say this strategy should be adopted". Two coders thought this message firstly disagreed with the previous message, and that the writer then proposed his/her opinion ("not an argument" with the supporting reasons [rest of the message]), and therefore, it should be labelled integration. However, the third coder thought it was just a personal opinion as suggestion for consideration without sufficient support, and labelled it exploration.

Second, messages with misleading language patterns could have impacted coders' decisions. Messages with language patterns, such as "consequently" or "both sides of" might indicate a conclusion or a "convergence" denoting the integration phase. However, such patterns could also be interpreted as a "leap to a conclusion" or a claim without supporting ideas, meaning that the messages should be classified in the phase of exploration. An example of such a message is, "Consequently, both sides of the arguments are equally compelling but have their share of fallacies. It depends on each person's confirmation bias to weigh a particular argument heavier than the other." These misleading language patterns tell us that some phrases and expressions can only be a possible predictor but not absolute evidence for classifying cognitive phases.

Most of the messages that were labelled as part of both the integration and resolution phases disputed whether the supporting ideas of new constructs were sufficient enough. This debate is very similar to the debates on distinguishing integration from exploration as discussed in this section.

Understanding the Reasons for the Disagreements

We found the bulk of the disagreements occurred between the exploration and integration phases. It may be because: (a) the proportion of messages in these two categories was much larger than in the other categories; (b) exploration and integration appear during the middle of a critical thinking activity, which tends to greater uncertainty, rather than at the beginning (awareness of a question) or the conclusion (outcomes after evaluation) stage; or (c) the criteria and instances of these two categories are ambiguous in the CoI framework, which is consistent with Rourke and Kanuka's (2009) critique about the lack of clear instances in Garrison et al.'s cognitive presence rubric. These reasons can also be connected with other critiques of the CoI. Garrison et al. (2001) borrowed from Dewey's (1933) five steps in reflective thinking to propose the four phases of cognitive presence. Still, they did not develop and elaborate on the theoretical foundations of Dewey's model (Jézégou, 2010). Garrison et al. (2001) merged the second ("diagnosis of a question") and third step ("suggestion of possible solution") from Dewey's (1933) model into the exploration phase. They renamed the fourth step ("elaboration of an idea by reasoning") as integration and the last step ("corroboration to form a concluding belief") as resolution. With respect to Dewey's model, the ambiguity of disagreements between exploration and integration occurred mainly when trying to distinguish "a suggestion of possible solution" (assigned to exploration) from the "elaboration of an idea by reasoning" (assigned to integration). In contrast, messages in the "diagnosis of a question step" (assigned to exploration) were easier for the coders to identify. Thus, we question whether the exploration phase should be separated back into the diagnosis step and a suggestion of possible solution step as defined in Dewey's (1933) model.

Henri and Lundgren-Cayrol (2005) proposed three phases of a collaborative learning approach for knowledge construction (e.g., exploration, elaboration, and evaluation), which intersect with the cognitive presence phases (Jézégou, 2010). The elaboration phase is positioned between exploration and evaluation, which is similar to the integration phase in cognitive presence. Henri and Lundgren-Cayrol (2005) also proposed two subcategories in the elaboration phase: *negotiation* and *validation*. The negotiation sub-phase refers to the learning processes that consider and collect other people's ideas to form diverse proposals of knowledge, and the validation sub-phase denotes consensus on the knowledge, reflecting multiple views (Henri & Lundgren-Cayrol, 2005). In this regard, the validation sub-phase is equivalent to the integration phase in cognitive presence. Interestingly, most of the ambiguous messages between

exploration and integration in our sample data could be assigned into a negotiation sub-phase. For example, one learner compared two opinions from previous comments and generated her/his own statements in a message, but the statements had not been supported by sufficient reasoning, which meant the message was more exploration and not yet integration. This example fits well into negotiation. Therefore, there may be a negotiation sub-phase between a "considerable solution" step (assigned to exploration) and a "consensus idea by reasoning" step (assigned to integration). This would be an additional phase in the cognitive presence schema.

Apart from ambiguities of language patterns in MOOC discussions and insufficiencies of the CoI framework, another significant factor that caused the disagreements between all the adjacent phases was the increase in the data sample size from a small scale to a vast magnitude. Using the taxonomies to categorise cognitive processes works well on a smaller scale. In comparison, the likelihood of outliers increases when researchers apply the taxonomies developed from a smaller-scale dataset to classify vastly larger samples (Mayer-Schönberger & Cukier, 2013). Also, online discussions have an informal, conversational flow that is relatively messy, and does not fit into the ordered phases in the CoI (Xin, 2012). We assume that investigating the general trend of cognitive processes within the messiness of communication in the myriad MOOC transcripts is more valuable than using rigid classification methods.

Categorising learners' discussion transcripts into single-label cognitive phases tends to be subjective and inaccurate. One possible solution is to label the MOOC discussion messages into multiple cognitive presence phases with confidence levels. Another solution would be to label the messages by multiple models of learners' critical discourse simultaneously. For example, Farrow et al.'s (2021) study applied both cognitive presence and the ICAP framework (Chi & Wylie, 2014). These methods could provide a richer portrait for the interpretation of learners' dialogue by different coders using different frameworks and would reflect the diverse variation in the discourse more authentically.

In response to our main research question (*Is our adapted coding rubric of cognitive presence a reliable tool to classify MOOC discussions?*), we conclude that although the adapted rubric of cognitive presence is a statistically reliable tool to classify the discussion messages from the LCT MOOC by three coders, an additional phase (negotiation) could be included to improve the rubric to accommodate the predominant disagreements between coders.

Limitations

We acknowledge the limitation that a classification rubric of cognitive presence developed for one discipline might not be generalisable to other domains. There are disciplinary differences in the expression of critical reflection and its assessment in the pedagogical designs of different courses. The evaluation of cognitive presence that is mainly based on textual information can be highly domain specific. We are aware that the research findings might only be reliable and valid for the classification of cognitive presence in our target MOOC, done by three coders.

Conclusion and Implications

This research offers several theoretical and practical implications. We have reported on a process of classifying cognitive presence using more coders and a larger dataset to cross validate an adapted coding rubric in a target MOOC. The overall result reveals good inter-rater reliability, indicating that the adapted rubric remains stable for classifying cognitive phases in MOOC discussions by more coders and with larger datasets. We have then dug deeper into the messages where coders disagreed between adjacent cognitive phases. The possible causes of the ambiguous categorisation between adjacent cognitive phases could be theoretical insufficiencies of the CoI, MOOC learners' informal writing styles, and the changes of data size in MOOCs. We envisage that negotiation may be the additional phase between exploration and integration where most disagreements occurred. Our findings can inform the ongoing refinement of the CoI framework and provide a foundation for an approach to developing automatic analysis of educator-learner dialogue at scale.

This study also has practical implications. For preparing the machine learning datasets, we suggest using multiple-label instead of single-label classification to analyse learners' cognitive presence in MOOC discussions. This takes into account learners' informal language usage in MOOC discussions. It provides learning analytics researchers with some hints for choosing algorithms and predictive features in the study of automatic cognitive analysis. For example, better prediction performance may be achieved using corpora that include both informal speech and formal writing texts to train and generate the numeric representations of discussion messages that are fed into machine learning algorithms (e.g., neural networks). Also, the computational linguistic tools (e.g., Coh-metrix), which were created to assess formal essay writing (McNamara et al., 2012), may not be appropriate for analysing MOOC discussion messages. The application of a reliable and smart automatic classifier for analysing the processes of critical discourse in open and distance learning at scale can potentially (a) enable learners to self-evaluate their learning, to complement the automatic learner grading systems, (b) be used to inform the design and adaption of course content, and (c) assist the assessment of educator-learner online dialogue efficiently in real time.

References

Alario-Hoyos, C., Estévez-Ayres, I., Pérez-Sanagustín, M., Kloos, C. D., & Fernández-Panadero, C. (2017). Understanding learners' motivation and learning strategies in MOOCs. *The International Review of Research in Open and Distributed Learning*, *18*(3), 119–137. https://doi.org/10.19173/irrodl.v18i3.2996

Amemado, D., & Manca, S. (2017). Learning from decades of online distance education: MOOCs and the community of inquiry framework. *Journal of E-Learning and Knowledge Society*, *13*(2), 21–32. http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/137/75

Barbosa, G., Camelo, R., Cavalcanti, A. P., Miranda, P., Mello, R. F., Kovanovic, V., & Gaševic, D. (2020). Towards automatic cross-language classification of cognitive presence in online discussions. In C. Rensing & H. Drachsler (Chairs), *LAK '20: Proceedings of the tenth international conference on learning analytics and knowledge* (pp. 605–614). ACM. https://doi.org/10.1145/3375462.3375496

Buchem, I., Amenduni, F., Poce, A., Michaescu, V., Andone, D., Tur, G., Urbina, S., & Šmitek, B. (2020). Integrating mini-moocs into study programs in higher education during COVID-19. Five pilot case studies in context of the open virtual mobility project. In *EDEN 2020 annual conference: Human and artificial intelligence for the society of the future* (pp. 299–310). https://doi.org/10.38069/edenconf-2020-ac0028

Cha, H., & So, H.-J. (2020). Integration of formal, non-formal and informal learning through MOOCs. In D. Burgos (Ed.), *Radical solutions and open science. Lecture notes in educational technology* (pp. 135–158). Springer Singapore. https://doi.org/10.1007/978-981-15-4276-3_9

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Corich, S., Kinshuk, & Hunt, L. (2006). Measuring critical thinking within discussion forums using a computerised content analysis tool. In S. Banks, V. Hodgson, C. Jones, B. Kemp, D. McConnell, & C. Smith (Eds.), *Proceedings of the 5th international conference on networked learning*, *2*(1), 1–8. http://www.lancaster.ac.uk/fss/organisations/netlc/past/nlc2006/abstracts/pdfs/P07 Corich.pdf

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process* (revised ed.). D. C. Heath.

Dillahunt, T., Wang, Z., & Teasley, S. D. (2014). Democratizing higher education: Exploring MOOC use among those who cannot afford a formal education. *The International Review of Research in*

Open and Distributed Learning, *15*(5), 177–196. https://doi.org/10.19173/irrodl.v15i5.1841

Farrow, E., Moore, J., & Gašević, D. (2021). A network analytic approach to integrating multiple quality measures for asynchronous online discussions. In M. Scheffel, N. Dowell, S. Joksimovic, & G. Siemens (Chairs), *LAK '21: Proceedings of the 11th international conference on learning analytics and knowledge* (pp. 248–258). ACM. https://doi.org/10.1145/3448139.3448163

Fleiss, J., Levin, B., & Paik, M. (2003). *Statistical methods for rates and proportions* (3rd ed.). John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/0471445428

Garrison, D. R. (2007). Online community of inquiry review: Social, cognitive, and teaching presence issues. *Journal of Asynchronous Learning Networks*, *11*(1), 61–72.

Garrison, D. R., & Anderson, T. (2011). *E-learning in the 21st century: A framework for research and practice* (2nd ed.). Routledge.

Garrison, D. R., Anderson, T., & Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education*, *2*(2), 87–105. https://doi.org/10.1016/S1096-7516(00)00016-6

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, *15*(1), 7–23. https://doi.org/10.1080/08923640109527071

Garrison, D. R., Anderson, T., & Archer, W. (2010). The first decade of the community of inquiry framework: A retrospective. *The Internet and Higher Education*, *13*(1–2), 5–9. https://doi.org/10.1016/j.iheduc.2009.10.003

Henri, F., & Lundgren-Cayrol, K. (2005). *Apprentissage collaboratif à distance: Pour comprendre et concevoir les environnements d'apprentissage virtuels* [Collaborative distance learning: Understanding and conceptualizing virtual learning environments]. Presses de l'Université du Québec.

Hu, Y., Donald, C., Giacaman, N., & Zhu, Z. (2020). Towards automated analysis of cognitive presence in MOOC discussions: a manual classification study. In C. Rensing & H. Drachsler (Chairs), *LAK '20*: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (pp.135–140). ACM. https://doi.org/10.1145/3375462.3375473

Jézégou, A. (2010). Community of inquiry in e-learning: A critical analysis of the Garrison and Anderson model. *The Journal of Distance Education / Revue de l'Éducation à Distance*, *24*(3), 1–18. http://www.ijede.ca/index.php/jde/article/view/707

Kanuka, H., Rourke, L., & Laflamme, E. (2007). The influence of instructional methods on the quality of online discussion. *British Journal of Educational Technology*, *38*(2), 260–271. https://doi.org/10.1111/j.1467-8535.2006.00620.x

Kaul, M., Aksela, M., & Wu, X. (2018). Dynamics of the community of inquiry (CoI) within a massive open online course (MOOC) for in-service teachers in environmental education. *Education Sciences*, *8*(2), Article 40. https://doi.org/10.3390/educsci8020040

Kovanović, V., Joksimović, S., Gašević, D., & Hatala, M. (2014). Automated cognitive presence detection in online discussion transcripts. In K. Yacef & H. Drachster (Eds.), *Workshop proceedings of LAK 2014.* Sun SITE Central Europe (CEUR). http://ceur-ws.org/Vol-1137/LA_machinelearning_submission_1.pdf

Kovanović, V., Joksimović, S., Poquet, O., Hennis, T., Čukić, I., De Vries, P., Hatala, M., Dawson, S., Siemens, G., & Gašević, D. (2018). Exploring communities of inquiry in massive open online courses. *Computers & Education*, *119*, 44–58. https://doi.org/10.1016/j.compedu.2017.11.010

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016). Towards automated content analysis of discussion transcripts. In D. Gašević & G. Lynch (Eds.), *LAK '16: Proceedings of the sixth international conference on learning analytics and knowledge* (pp. 15–24). ACM. https://doi.org/10.1145/2883851.2883950

Liu, C.-J., & Yang, S. C. (2014). Using the community of inquiry model to investigate students' knowledge construction in asynchronous online discussions. *Journal of Educational Computing Research*, *51*(3), 327–354. https://doi.org/10.2190/EC.51.3.d

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McKlin, T. (2004). *Analyzing cognitive presence in online courses using an artificial neural network* [Doctoral dissertation, Georgia State University]. ScholarWorks @ Georgia State University. https://scholarworks.gsu.edu/msit_diss/1/

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Mladenić, D. (2010). Feature selection in text mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 406–410). Springer US. https://doi.org/10.1007/978-0-387-30164-8_307

Neto, V., Rolim, V., Ferreira, R., Kovanović, V., Gašević, D., Dueire Lins, R., & Lins, R. (2018). Automated analysis of cognitive presence in online discussions written in Portuguese. In V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink, & M. Scheffel (Eds.), *EC-TEL 2018: Lifelong technology-enhanced learning. Lecture notes in computer science, vol. 11082* (pp. 245–261). Springer International. https://doi.org/10.1007/978-3-319-98572-5_19

Park, C. L. (2009). Replicating the use of a cognitive presence measurement tool. *Journal of Interactive Online Learning*, *8*(2), 140–155. http://www.ncolr.org/jiol/issues/pdf/8.2.3.pdf

Phan, T., McNeil, S. G., & Robin, B. R. (2016). Students' patterns of engagement and course performance in a massive open online course. *Computers and Education*, *95*, 36–44. https://doi.org/10.1016/j.compedu.2015.11.015

Rourke, L., & Kanuka, H. (2009). Learning in communities of inquiry: A review of the literature. *Journal of Distance Education*, *23*(1), 19–48. http://www.ijede.ca/index.php/jde/article/view/474/815

Sadaf, A., & Olesova, L. (2017). Enhancing cognitive presence in online case discussions with questions based on the practical inquiry model. *American Journal of Distance Education*, *31*(1), 56–69. https://doi.org/10.1080/08923647.2017.1267525

Siemens, G. (2013). Massive open online courses: Innovation in education? In R. McGreal, W. Kinuthia, & S. Marshall (Eds.), *Open educational resources: Innovation, research and practice* (pp. 5–16). Commonwealth of Learning and Athabasca University. https://www.oerknowledgecloud.org/archive/pub_PS_OER-IRP_web.pdf

Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, *29*(2), 217–257. https://doi.org/10.1007/s40593-019-00174-2

University of Auckland. (n.d.) *Logical and critical thinking*. FutureLearn. https://www.futurelearn.com/courses/logical-and-critical-thinking

Xin, C. (2012). A critique of the community of inquiry framework. *The Journal of Distance Education*, *26*(1), 1–7. http://www.ijede.ca/index.php/jde/article/download/755/1333?inline=1