Received: 13 November 2019 | Accepted: 19 July 2020

DOI: 10.1002/cae.22314

SPECIAL ISSUE ARTICLE

WILEY

Is critical thinking happening? Testing content analysis schemes applied to MOOC discussion forums

Tim O'Riordan[1] | **David E. Millard**[2] | **John Schulz**[3]

[1]Web and Internet Science Research Group, University of Southampton, Highfield, Southampton, UK

[2]Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton, UK

[3]School of Education, University of Southampton, Highfield, Southampton, UK

Correspondence
Tim O'Riordan, Web and Internet Science Research Group, University of Southampton, Highfield, Southampton SO17 IBJ, UK.
Email: tjor1@yahoo.com

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: 1383089

Abstract

Learners' progress within computer-supported collaborative learning environments is typically measured via analysis and interpretation of quantitative web interaction measures. However, the usefulness of these "proxies for learning" is questioned as they do not necessarily reflect critical thinking—an essential component of collaborative learning. Research indicates that pedagogical content analysis schemes have value in measuring critical discourse in small scale, formal, online learning environments, but research using these methods on high volume, informal, Massive Open Online Course (MOOC) forums is less common. The challenge in this setting is to develop valid and reliable indicators that operate successfully at scale. In this study, we test two established coding schemes used for the pedagogical content analysis of online discussions in a large-scale review of MOOC comment data. Pedagogical Scores are derived from manual ratings applied to comments by raters and correlated with automatically derived linguistic and interaction indicators. Results show that the content analysis methods are reliable, and are very strongly correlated with each other, suggesting that their specific format is not significant in this setting. In addition, the methods are strongly associated with the relevant linguistic indicators of higher levels of learning and have weaker correlations with other linguistic and interaction metrics. This suggests promise for further research using Machine Learning techniques, with the goal of providing realistic feedback to instructors, learners, and learning designers.

KEYWORDS

content analysis, discussion forums, education, MOOC

1 | INTRODUCTION

Online discussions within computer-supported collaborative learning (CSCL) environments are a useful source of data for those involved in analysing educational interventions, as they reflect engagement with learning and offer a "gold mine of information concerning...the acquisition of knowledge and skills" [30, p. 118].

*Tim O'Riordan, Independent Researcher, 117 Newton Road, Southampton, SO18 1NH, United Kingdom

Dialogue helps learners' build personal social capital and gain exposure to new ideas [41], and language used in these environments has been shown to indicate the depth of critical thinking [1]. However, educational research has predominately focused on discussion within formal learning environments rather than the informal setting provided by Massive Open Online Courses (MOOCs) [88] and has been dominated by assessments of the quantity rather than the quality of interaction.

Discussion forums within online learning environments have been identified as rich seams of instructor and learner interaction data that can be mined to monitor levels of participation, but they can also reveal significant aspects regarding the quality of interaction through the adoption of content analysis techniques [11,33,50]. Weber defines these techniques as research methods that build on "procedures to make valid inference[s] from text" [84, p. 1], and in this study, we are proposing to interpret and categorise critical thinking within MOOC comment forums (e.g., Reference [25]).

Although a precise definition of critical thinking is unresolved, it is recognised as a key objective when encouraging learners to adopt in-depth, rather than surface, learning approaches [55]. In the context of this study, we agree with Lipman who argues that critical thinking is best acquired within the social context of a community of inquiry [49], as well as Biggs [5] association of deep learning with "affective involvement" through interaction.

Dialogue between learners stimulates cognitive conflict which encourages reflection, assimilation of new knowledge, and continued interaction. In this context "critical thinking" is perhaps best defined as "reasonable and reflective thinking that is focused upon deciding what to do or believe" [56, p. 1].

Although the high volume of MOOC data provides an unprecedented opportunity for insight into how CSCL is used in practice, the reliability of analysis methods used to explore this data is questioned. Specifically, some argue that the methods lack coherence and validity [15], and others identify a research-inhibiting lack of consistency in their application [85].

A high volume of data combined with uncertainty regarding analysis methods emphasises the importance of constructing theoretically sound methods that can reliably, and automatically, analyse this data. To address these issues, this study employs appropriate pedagogical content analysis methods, using instruments that have previously been adopted in studies exploring the depth of critical thinking evidenced in CSCL, and seeks to explore the potential of established methods for identifying critical thinking in MOOC forums.

In particular, our study aims to identify features that would distinguish behaviour suggestive of levels of critical thinking, and sets out to answer three questions:

- RQ1: Are coding schemes used for pedagogical content analysis of online discussions reliable in the context of MOOC discussion forums? In particular, can different people consistently apply them, and do different frameworks identify the same levels of critical thinking?
- RQ2: Are the linguistic characteristics of comments significant indicators of levels of critical thinking when applied to MOOC discussion forum comments, as identified through pedagogical content analysis?
- RQ3: To what extent do more typical measures of attention to learning (such as social interactions) indicate levels of critical thinking when applied to MOOC discussion forum comments, as identified through pedagogical content analysis?

Our work lays the foundations for further research into the analysis and visualisation of Web-based learning, with the potential to improve learner reflection, MOOC development tools, and the discoverability of high-quality learning materials.

2 | BACKGROUND

As the ultimate aim of this study is to develop automated methods of assessing comments that can be readily comprehended by users (i.e., educators), it is important to ground this method on established pedagogic theory. In Weltzer-Ward's [85] analysis of 56 content analysis-coding schemes used between 2002 and 2010, Bloom's Taxonomy [7], and analyses adopting Community of Inquiry: Cognitive presence (CoI) [25] were recognised as established methods with high citation counts, accounting for a high number of reviewed papers. They are, therefore, a good choice for the content analysis in our study and inform the rubrics we developed which follow in the tradition of similar work that adopts these methods. In this section, we will review these manual coding schemes and give an overview of existing work in the linguistic and interaction analysis.

2.1 | Bloom's Taxonomy of the cognitive domain

The Taxonomy of Educational Objectives: Handbook 1: Cognitive Domain, commonly referred to as "Bloom's Taxonomy" was developed to improve the "exchange of ideas and materials among test workers, as well as other persons concerned with educational research" [6,7, p. 1], and promote the use of teaching methods that encourage higher-order learning. Though directed by a small committee, the Taxonomy resulted from a collaborative effort that took input

and feedback from a wide range of educators, educational psychologists, administrators and researchers.

Bloom's Taxonomy consists of a hierarchy of categories of educational goals or outcomes, starting from the lower-order learning goals of "remember" and "understand", to the mid-level uses of knowledge as evidenced in "apply" and "analyse", with "evaluate" and "create" indicating the achievement of deeper understanding. Further Taxonomies classifying the Affective and Psycho-Motor domains were published but did not reach the high level of recognition and use achieved by Handbook 1.

Researchers have found the categories useful as a framework for analysing learning processes. Bloom himself used them to evaluate types of learning that take place in class discussions and compared them with lectures [6]. His key finding that learners spend more time engaged in higher-order thinking in class discussion than in lectures, led him to suggest that increasing opportunities for learner interaction would lead to improved development of problem-solving skills.

Kember's [37] association of Bloom's dimensions with Mezirow's [51] "thoughtful action" category (e.g. writing), Gibson, Kitto, and Willis' [28] use of Bloom to map word types to levels of cognition, and Karaksha et al.'s [35] use of Bloom to evaluate the impact of e-learning tools in a higher education setting, support the use of the Taxonomy in this study.

Furthermore, in Chan et al.'s [11] study two raters were employed to analyse essay papers and classroom discussions, applying Bloom, Structure of the Observed Learning Outcomes (SOLO), and the reflective thinking measurement model. Finding strong correlations between the models in long essays, but not in short discussions, they proposed further research using more than two raters to improve agreement and using the new version of Bloom to improve the accuracy of assessing cognitive learning outcomes. By engaging a team of seven raters and Krathwohl's updated version of Bloom [42] in this study, we aim to advance understanding of these research issues.

Table 1 presents the categories used by human raters in this study and includes descriptions and verb types associated with each category.

2.2 | Community of Inquiry

Community of Inquiry is based on the interaction of the forms of engagement or "presence" within Web-based learning communities: Cognitive presence, social presence, and teaching presence [26]. As our study looks for evidence of critical thinking in MOOC forums, our focus is on the cognitive presence dimension, which Garrison, Anderson, and Archer [25] define as "critical, practical inquiry" as evidenced within four types of text-based dialogue: Triggering, exploration, integration, and resolution [57, p. 14]. These categories refer to stages of dialogue—starting with an initiating "triggering" comment and ending with assertions that conclude the discussion (Table 2).

As an established pedagogic content analysis method, CoI has been applied within many studies. In her paper exploring the application of learning analytics, Dringus asserts that CoI provides "an array of meaningful and measurable qualities of productive learning and communication in online learning environments", and suggests converting CoI dimensions into datatypes that can

TABLE 1 Bloom's Taxonomy [7,11,42]

Score	Descriptor
0—Off-topic	There is written content, but it is not relevant to the subject under discussion.
1—Remember	Recall of specific learned content, including facts, methods, and theories. Verbs: Name, describe, relate, find, list, write, tell.
2—Understand	Perception of meaning and being able to make use of knowledge, without understanding full implications. Verbs: Explain, compare, discuss, restate, predict, translate, outline.
3—Apply	Tangible application of learned material in new settings. Verbs: Show, complete, use, classify, examine, identify, investigate, categorise, differentiate, organise.
4—Analyse	Deconstruct learned content into its constituent elements to clarify concepts and relationships between ideas. Verbs: Explain, compare, contrast examine, illustrate, implement, solve.
5—Evaluate	Assess the significance of material and value in specific settings. Verbs: Check, decide, rate, choose, recommend, justify, assess, prioritise, critique.
6—Create	Judge the usefulness of different parts of content and producing a new arrangement. Verbs: Synthesise, invent, plan, compose, construct, design, imagine, generate.

TABLE 2 Community of Inquiry: Cognitive presence [26,57]

Score	Descriptor
0—Off-topic	There is written content, but it is not relevant to the subject under discussion.
1—Triggering event	A contribution that exhibits a sense of puzzlement deriving from an issue, dilemma or problem. Includes contributions that present background information, asks questions or moves the discussion in a new direction. Verbs: Evoke, induce, contradict.
2—Exploration	A comment that is seeking a fuller explanation of relevant information. This can include brainstorming, questioning and exchanging information. Contributions are unstructured and may include: Unsubstantiated contradictions of previous contributions, different unsupported ideas or themes, personal stories and descriptions or facts that are not used as evidence. Verbs: Inquire, diverge, search.
3—Integration	Previously developed ideas are connected. Contributions include references to previous messages followed by substantiated agreements or disagreements, developing and justifying established themes, cautious hypothesis, combining different sources, providing a tentative solution to an issue. Verbs: Test, conjecture, check.
4—Resolution	New ideas are applied, tested and defended with real-world examples. This involves methodically testing hypotheses, critiquing content in a systematic manner and expressing supported intuition and insight. Verb: Commit, settle, confirm.

be mined to "draw out coherent patterns" [19, p. 96] in online courses.

Tirado, Hernando, and Aguaded [78] apply CoI in their study on the quality of knowledge construction in social environments and call for the strong validation of content analysis methods that evaluate the processes of the construction of knowledge in this setting. Shea et al. [69] adapt the approach to measure the students' practice of successful learning strategies and compare their results with social network analysis methods. They recognise the importance of further research into the relationship between cognitive presence and interaction and suggest that its detection contributes to the understanding of learners' networking behaviours. Joksimovic et al. [33] associate linguistic proxies for learning with CoI stages in discussion forums within small scale online courses. Their findings indicate the usefulness of further research that explores the effects of different levels of cognitive presence on learners with different levels of prior knowledge.

Finally, Waters et al. [83] implement a machine learning approach to predict students' critical thinking levels in formal online discussions according to CoI. In their study, they adopt word count, post similarity, chronological order, and other features to build a model that achieves a moderate level of accuracy. Although we do not implement an automated approach, our study adopts Waters et al.'s suggestions for future work that include the use of Linguistic inquiry and word count (LIWC) analysis to identify phases of critical thinking.

The research questions identified by these studies consider the utility of data derived from pedagogical content analysis and its potential in measuring performance in online courses. These are relevant to our study, which seeks to contribute to the development of robust and effective ways of understanding large-scale comment data that are based on established theory.

2.3 | Bloom and CoI as content analysis methods

Weltzer-Ward [85] argues that understanding of online environments may be improved through the use of pedagogical content analysis methods, and calls for research on their application outside of online academic classroom contexts, and the exploration of opportunities for synthesis. Our study addresses these research areas by using different methods that adopt complementary approaches to analyse discussion in large scale, informal settings. Although Garrison et al. [25] acknowledge the consistency of their framework with socioconstructivist learning theory emerging from Dewey's [17] ideas on the importance of sociological as well as psychological aspects of learning, Bloom et al., do not explicitly recognise a single theoretical basis. However, the authors of these frameworks adopt hierarchical approaches identifying changes in learners' behaviour that have much in common with Piaget's theory of staged development [64] and

implicitly recognise the value of social learning explored by Vygotsky [81].

Although there are similarities, there are also distinct differences in their focus as well as approaches to the evaluation of critical thinking: Bloom facilitating generalisable evaluations of educational outcomes that can be applied to assessing learners in any number of settings, and CoI focusing on the appraisal of participation in the specific CSCL environment. In addition, some educational psychologists argue that individual and distributed cognition are two distinct, interrelated processes [53], and the methods we adopt in this study emphasise these different aspects: Bloom—individual, and CoI—distributed. By comparing two distinct approaches to measure critical thinking, we seek to establish if different methods yield significantly different results and identify opportunities for synthesis. This leads to our first hypothesis (RQ1) that there are significant differences between levels of critical thinking as measured by each method.

In this study, we evaluate MOOC forum comments in terms of the extent to which they provide evidence of deep learning approaches that reflect critical thinking through the lens of each method and attempt to identify significant differences in outcome from their use. Specifically, we seek to establish if these pedagogical content analysis methods can be applied consistently by different people and if these methods identify the same types of learning activity.

In addition to comparing and critically evaluating two different pedagogical content analysis methods, we compare the outcomes of this analysis with established proxies for learning in the form of linguistic analysis and typical interaction analysis, these are each explored in the next two sections.

2.4 | Linguistic analysis

The content and style of language used in everyday communications provide important indicators of psychological and social meaning that may be measured by quantitative methods, including content analysis and word pattern analysis [62]. Characteristic approaches to quantitative language analysis involve the identification and coding of similar patterns and the interpretation of content supported by statistical tests of significance [39]. Writing, speech, and the types of words used are seen as important proxies for emotional and cognitive processes [60].

In recent years, increased emphasis on content analysis studies exploring language use and CSCL has been identified [85]. For example, Delfino and Manca [16] discuss the use of "figurative" language in online social contexts, Miller [52] explores gender-related language patterns, and Uzuner [79] identifies educationally valuable talk in Computer-Supported Collaborative Learning (CSCL), Tausczik and Pennebaker [76] adopt a real-time language feedback system to improve learner collaboration, Robinson, Navea, and Ickes [65] correlate student language use with educational attainment, Joksimovic et al. [33] correlate word categories with CoI dimensions, and Allen, Snow, and Mcnamara [1] use linguistic indicators to predict learners' reading comprehension abilities. Evidence that pedagogically meaningful dialogue in Web-based environments can be automatically identified using learning analytic techniques has importance for this study [14], as does the use of mixed linguistic and interactional data to identify potentially "at-risk" learners [86].

2.5 | Linguistic inquiry and word count

Among computational approaches to language analysis, LIWC [24] was chosen as suitable for analysis of online discussion, and evaluation of cognitive processes. LIWC was developed as a result of studies into the therapeutic effects of writing about psychological traumas. The application adopts a quantitative, word count approach that aims to reveal the psychological meaning of words taken out of context from their original settings [62]. It searches within text files for over 2,300 words or word stems, tracking stylistic aspects of language use classified into 82 dimensions (e.g., articles, prepositions, pronouns), psychological processes (e.g., positive and negative emotion categories, cognitive processes), and other categories.

In addition to the developers' experiments aimed at validating the program, there are a number of studies which suggest the usefulness of LIWC in detecting the meaning of words. Although not seen as a replacement for qualitative analysis, Carroll [9] found LIWC provided meaningful results in their analysis of essays written for a critical thinking course. He discovered learners demonstrated less use of pronouns and words related to insight (think, know, consider), discrepancy (should, would, could), and tentativeness (maybe, perhaps, and guess), and were more likely to express causal thinking (because, effect, hence) in their final essays, compared with writing at the start of the course.

2.6 | LIWC categories

In this study, though we explored correlations with all LIWC categories, previous research indicates that specific

categories are more closely associated with critical thinking than others. Therefore, as we set out to answer the question of whether the linguistic characteristics of comments are reliable proxies for levels of critical thinking (RQ2), we expected to find significant associations between the results of our pedagogical analysis and several LIWC characteristics.

2.7 | Word count

The number of words used in comments is often understood as a rough guide to levels of participation [3] and is commonly associated with the intensity of engagement [13,68]. Ferguson and Buckingham Shum's [21] research into synchronous text chat, and Joksimovic et al.'s [32] linguistic analysis of online discussions similarly suggest close associations between high word counts and thoughtful, "exploratory" exchanges.

2.8 | Pronouns

In their comparison of self-assessment with traditional (nonreflective) assignments, Peden and Carroll [58] found that learners writing self-assessment essays included more pronouns, insight and emotion words and used simpler language than expressed in traditional academic assignments. Kacewicz et al. [34] suggest that higher status contributors use fewer first-person singular pronouns, and Vosecky, Leung, and Ng's [80] research into tweet quality suggests that "I-talk" signifies "low quality", nonfactual communication. Robinson, Navea and Ickes [65] discovered that they could predict learners' course performance on their use of "word simplicity, first-person singular pronouns, present tense, details concerning home and social life, and words pertaining to eating, drinking, and sex" (p. 469), concluding that low-performing learners tended to exhibit egocentricity in their writing.

2.9 | Causal words

Within LIWC dictionaries, causal words are categorised as a subgroup of cognitive process words, which suggest an engagement with active reappraisal, or processing of information [61]. Although Joksimovic et al. [33] found counts of causal words were not significant between higher phases of CoI, several studies (e.g., References [34,46,59]) have found that causal words are related to the level of cognition. Linguistic analysis of journals and essays indicates that causal words are more evident in precise and concise descriptions, and indicate progress in the level of cognition and understanding [59]. In addition, increased levels of differentiating between competing ideas have been linked to higher levels of cognition [76].

2.10 | Power and affiliation

The LIWC categories of power and affiliation, are developed from thematic apperception test (TAT) research and relate to assessments of an individuals' unconscious drives and social motives, where the affiliation motive is related to the friendliness and establishing rapport, and power is associated with making an impact and exerting control [87]. Although the literature does not suggest causal associations between TAT scores and levels of critical thinking, in LIWC higher incidence of power words suggests the writers' perception of themselves as having high status or expertise. In this study, we conceptualise this form of self-assurance as a potential indicator of critical thinking.

2.11 | Emotion words

Using sentiment analysis to measure relationships between mood and different variables—from consumer confidence to managing disaster relief—is commonplace wherever people's behaviour is under scrutiny. Research suggests that though positive language can suggest a focus on group cohesion, which may encourage individuals to work harder [23,47], it has been noted that correlation with positive sentiment can suggest disconnection, and that high levels of empathetic discussion may distract learners from key tasks [47]. Conversely, the expression of negative sentiment has been associated with "cognitive disequilibrium" and higher levels of thinking [18,27].

2.12 | Word length

Although complex cognitive processes and critical thinking are often associated with using long words [3,38], some researchers have found that counts of long words are not significant indicators of cognitive load, but have use in supporting analysis that includes other significant features [38]. However, long sentences should not necessarily suggest increased cognitive attention. In their research into predictors of students' reading comprehension, Allen et al. [2] assert that shorter sentences can suggest more sophisticated writing strategies.

2.13 | Other word types

Researchers have found negation, auxiliary verbs, and conjunctions to be significant indicators of cognitive load [39], and in the analysis of undergraduate writing, these categories have shown significant differences between triggering and other phases of CoI [33]. Research also indicates a high incidence of prepositions associated with attention to reflective behaviour. High use of prepositions are identified as significant indicators of increased cognitive load [38], and their prevalence in the discussion sections of journal articles, which are "often the most complex part of an article" [31, p. 35].

In addition, Joksimovic et al.'s [33] study found distinct use of the dictionary, functional, inhibition, inclusive and cognitive words, as well as articles, prepositions, conjunctions in the triggering phase, but found no significant difference in the use of pronouns or insight words throughout the four phases.

2.14 | Limitations

Although supported by numerous research outputs, linguistic analysis is limited in its reliability [75]. In addition to uncertainty over the meaning of higher numbers of words per sentence counts, as referred to above, analysis of word categories may also be compromised. Contributors to discussion forums often use symbolic, oblique and indirect ways of communicating meaning, which may lead to classification errors [60]. Multiple meanings of words, complicated sentence formation, and unclear use of pronouns may obscure meaning and require more complex methods to resolve uncertainty than are available in the software used in this study [31]. However, notwithstanding the potential for error, we agree with Pennebaker and Francis' claim that LIWC analysis is "as valid as a judge-based system that requires multiple judges who, themselves, are prone to error" [39, p. 622].

2.15 | Interaction analysis

Where use of language acts as a more-or-less unconscious indicator of mood, interaction analysis looks at more direct actions. "Likes" are a common intentional rating mechanism used to signify personal feelings [45,77]. Some research suggests that this metric is ambiguous [43] and unreliable [74], however, this indicator, as well as sentiment analysis, are widely used in learning analytics, for example, to identify learner attrition [86], self-confidence [71], and learners' opinions of courseware [74]. In addition, there is some evidence that this cumulative rating system may provide learners with timely prompts that can lead to higher levels of learning [13], and the platform hosting the MOOCs explored in this study include a "like" button feature associated with all individual comments, which allows learners to provide immediate, simple feedback.

Furthermore, by placing discussion forums within the context of each activity and providing mentor support [48], the platform encourages sharing and situated debate [4,36], with the explicit intent of building communities of inquiry and inspiring higher-level learning. In this context, we set out to discover if learners' use of the "like" button was significantly associated with pedagogical content analysis methods. Specifically, we aimed to answer the question of whether the number of likes awarded to comments or the sentiment of posts are a reliable indicator of the level of critical thinking (RQ3).

3 | METHODOLOGY

To answer our three research questions the comment data from three Massive Open Online Courses (MOOCs) offered on the FutureLearn platform in 2014–15 were analysed. The MOOCs were chosen to facilitate the analysis of writing produced in diverse subject areas: business, education, and science. More than 41,500 registered learners engaged with the courses, with nearly 15,000 contributors posting over 174,500 comments containing more than 8.5 million words (the MOOC2015 corpus). Each MOOC was delivered via an average of 20 "steps" per week throughout each of the 3- to 6-week courses, and each step provided the facility for instructors and registered learners to contribute to discussions within the steps' comment field. As all comment data was provided in anonymised form, it was not possible to separate comments by type of participant. Although the random sample of 1,500 comments used in this study may have included comments from instructors as well as learners, as the literature reports low levels of instructor intervention (e.g., Reference [8]) our assumption is that the bulk of comments were made by learners.

Sample size was limited by the time each rater would have to provide an reliable evaluation, the cost of employing raters, and the available financial resources. To obtain accurate results, I anticipated that raters had to be motivated to undertake the tasks in an expert manner. Many studies support the proposition that the strongest motivating factor for this type of work is the ability to earn money [8,36]. As raters were expected to carry out expert assessments, we decided to pay them the normal rate for similar professional activity (e.g., teaching).

Having undertaken similar work in earlier studies, we estimated that each rater would expect to spend an average of 30 s evaluating each comment. As each comment was rated twice (once for each analysis method), we estimated that within our small research budget raters could be reasonably expected to evaluate no more than 1,500 comments in total (the MOOC2015 corpus).

Although amounting to less than 1% of the total number of comments submitted on the three chosen MOOCs, the MOOC2015 corpus is considerably larger than 20–140 sample sizes required to train "good classifiers" suggested by Beleites et al.'s [4] algorithm design research.

To select 1,500 comments from a total of 174,500, we used a simple random sampling method [67]. Labelling each comment with a unique number and choosing 500 comments from each MOOC using a random number generator would have provided a satisfactorily random sample. However, raters expressed concern during training that comments selected in this way would be seen out of context and may have led to inaccurate ratings. As individual comments taken out of the context could be misconstrued by raters, we organised them into batches of 20 consecutive comments (the minimum number of comments considered to be large enough to facilitate context-based judgements). Eight batches from each MOOC were then selected for rating using a random number generator [29]. Three randomly selected batches of 20 consecutive comments from each MOOC were also selected to facilitate test rating, before undertaking analysis of the rest of the sample (see Figure 1).

Qualitative analysis was undertaken by a team of seven raters recruited from postgraduate students registered at a UK university, using content analysis methods based on Bloom's Taxonomy (Table 1) and CoI (Table 2), to rate whole comments (Figure 2). Two of the seven had backgrounds in education, two in anthropology, and one each from physics, psychology, and languages. Five had previous experience of assessing written work. The raters were provided with a short face-to-face instruction session, where they scored a variety of typical comments and observed how others scored the same comments. They were instructed to work alone on the coding task and not compare results with, or request advice from, anyone.

To identify outlying scores and possible misunderstandings among raters, an initial test selection of 60 comments, comprising 20 randomly selected consecutive comments from each MOOC, were scored by the coding team. Intraclass correlation coefficients were calculated using a two-way mixed, consistency, average measures definition and provided inter-rater reliability scores of 0.93 for Bloom and 0.898 for CoI suggesting "almost perfect" agreement [67, p. 165], and indicating that levels of critical thinking were scored similarly across raters.

Rating of a larger sample then went ahead. Comments from all three MOOCs were numbered in batches of 60 consecutive comments from which eight batches were randomly chosen and distributed among the raters—two batches from each MOOC. Raters scored each comment twice (once by Bloom and once by CoI), with each batch being scored by two raters working independently. To avoid confusion between content analysis methods, scoring sheets were distributed with a 10-day time lag between each method, and only after the scoring using the first method had been completed. In total 1,440 comments were scored.

Comments from the test and full sample were combined ($n = 1{,}500$) and Pedagogical Scores (PS) for each comment were generated based on the average score of the two raters who had examined that comment, where the PS value is equivalent to the level identified for that comment within the analysis framework.

Statistical analysis software was used to conduct two-way analyses of variance and generate scatter plots with fitted lines to identify the existence and intensity of simple linear regression. Histograms show close to normal distribution of mean scores (Figures 3 and 4, and Table 3) and these scores were compared with number of words per comment, number of likes, and LIWC2015 categories to produce correlation and prediction values. LIWC2015 word category analysis is believed to be unreliable for texts containing less than 50 words [63], and as 40% of comments contained less than this amount, results were explored on three levels: 1,500 individual comments (where analysis methods, likes, and word count were compared), 150 aggregated batches of 10 contiguous comments (LIWC2015 compared with average scores for each batch), and 607 individual comments containing 50 or more words (LIWC2015 compared with individual average scores). The aggregated batches were selected from the data sets by grouping together the text from 10 contiguous comments. Average PS was calculated for each block and correlated with LIWC2015 analysis.

In addition, PS generated by the two different frameworks (CoI and Bloom) were correlated to explore whether the frameworks were measuring the same sorts of pedagogical activity.

4 | RESULTS

Table 4 shows the results generated when comparing the pedagogical analysis methods with each other, with

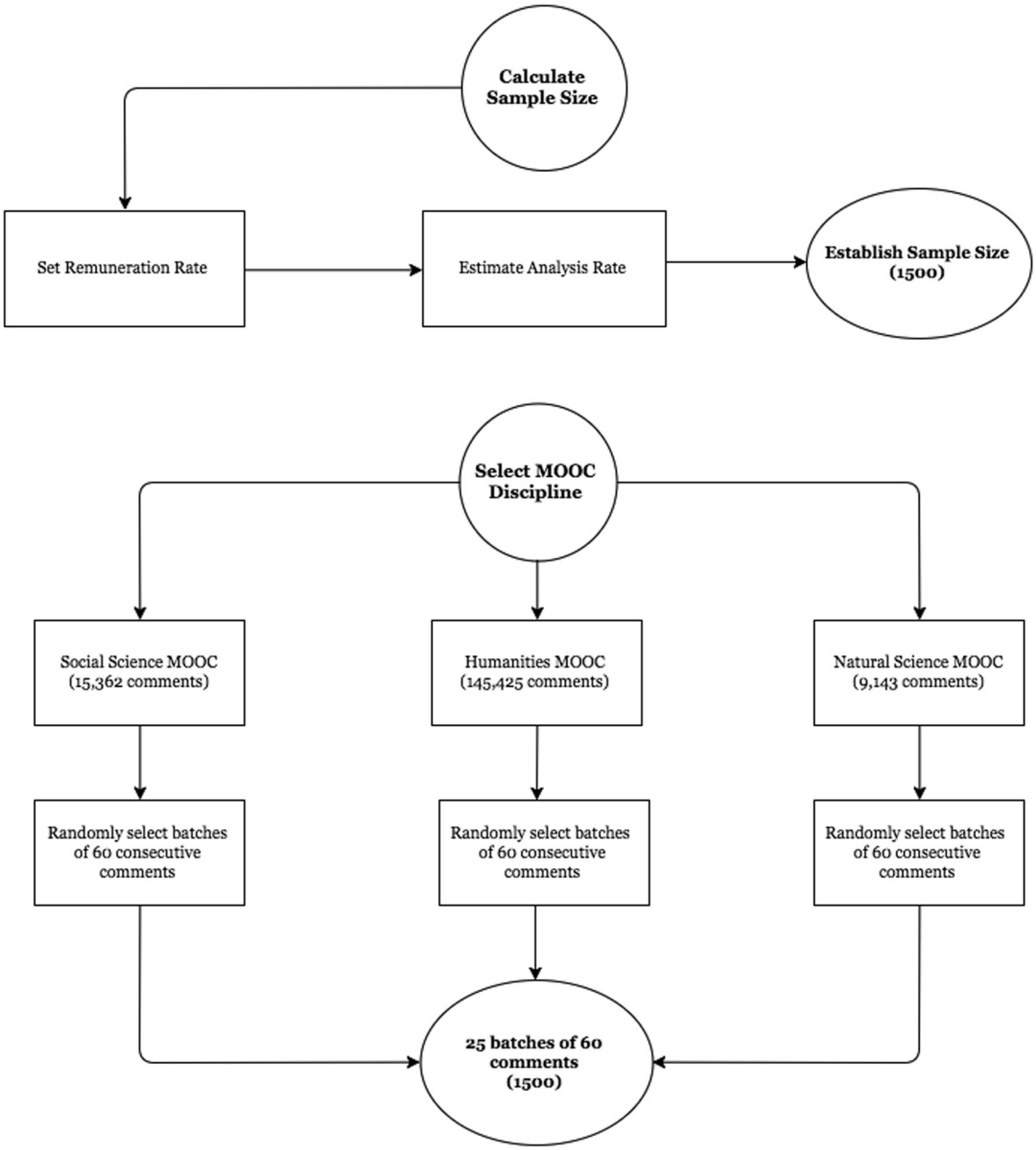

FIGURE 1 Sample selection process. MOOC, Massive Open Online Course

learner attention (in the form of likes), and with a wide range of linguistic analysis metrics. When classifying the strength of correlations, this study follows Evans [20], who classifies *r* values of less than .20 as very weak, .20 to .39 as weak, .40 to .59 as moderate, .60 to .79 as strong, and .80 or greater as very strong correlations. In Table 4, we have shaded the cells to show the two strongest, significant correlations: word count and first-person pronoun count. Due to the number of significance tests (a total of 103 tests) we have made note of *p* values at <.05, <.01, and <.00.

4.1 | Research Question 1: The reliability of the pedagogical analysis methods

As the key test of objectivity in content analysis research, establishing the degree to which raters agree is vital, unfortunately, many studies either fail to report rater agreement, or report discussion leading to full agreement [15]. Krippendorff [43] argues that this approach is of little use as it tests just the reliability of the individual raters rather than the method. As there is no settled

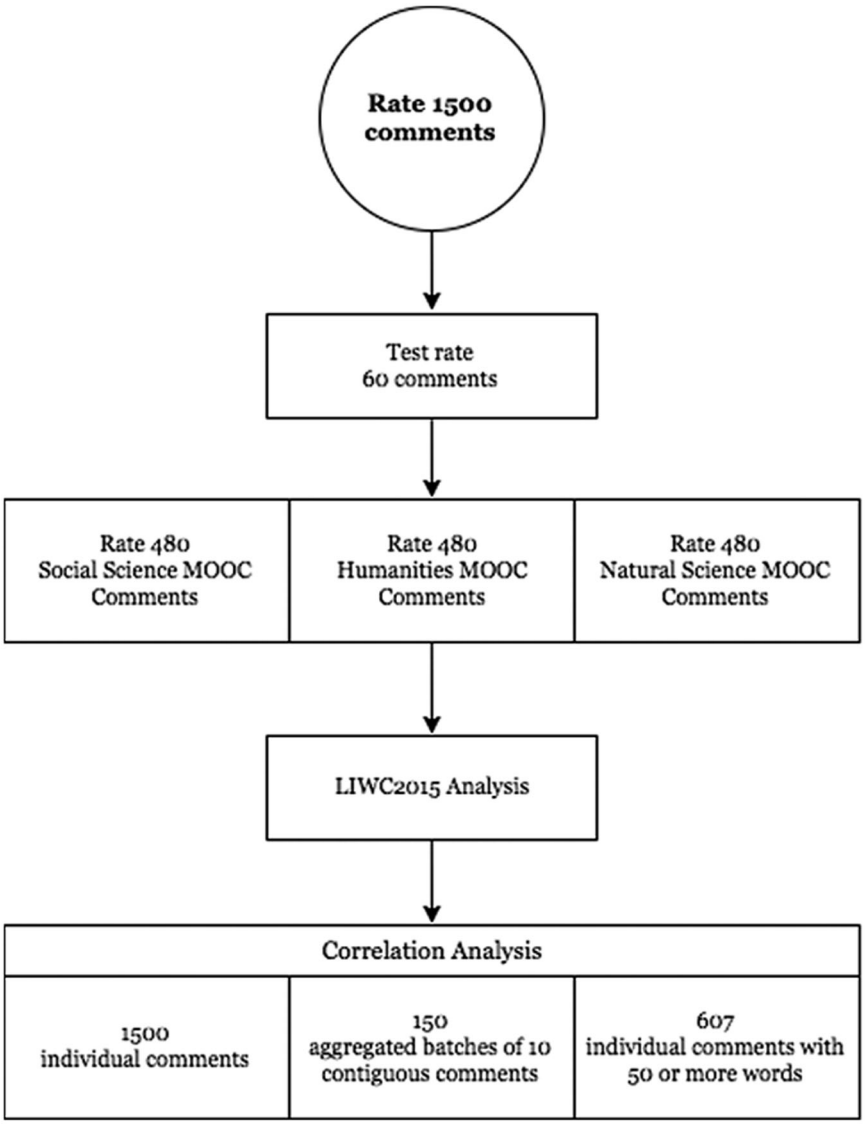

FIGURE 2 Comment rating process. LIWC, linguistic inquiry and word count; MOOC, Massive Open Online Course

method of testing agreement, our study follows de Wever et al.'s [15] recommendation and reports two indices.

To establish the reliability of the pedagogical analysis methods used in this study, intraclass correlation coefficients were calculated between pairs of raters and provided inter-rater reliability scores of 0.832 for Bloom and 0.818 for CoI. This suggests a high degree of agreement between raters and indicates that the pedagogical frameworks were interpreted and applied similarly across raters. Furthermore, reliability analysis using the Krippendorff's [44] α method provided inter-rater reliability scores of 0.7287 for Bloom and 0.6961 for CoI, which supports the use of these methods to reach tentative conclusions.

When comparing PS derived from the two frameworks there is a high correlation score of .909 ($p < .001$), suggesting a close association between Bloom's levels of learning and CoI's measures of meaningful and productive discourse. This suggests that while they describe pedagogical activity in different ways, they are relatively consistent in measuring its presence and strength.

4.2 | Research Question 2: Linguistic content analysis as an indicator of learning

We sought to establish which LIWC characteristics had significant correlations with levels of critical thinking using the LIWC2015 software to examine all characteristics. Two moderate to strong indicators were identified across all approaches to corpus analysis: word count (Figure 5; highest correlation: $r = .759$, $p < .001$) and first-person singular pronouns (Figure 6; $r = -.533$, $p < .001$).

Moderate indicators in aggregated comments that were also weak indicators in ≥50-word comments were

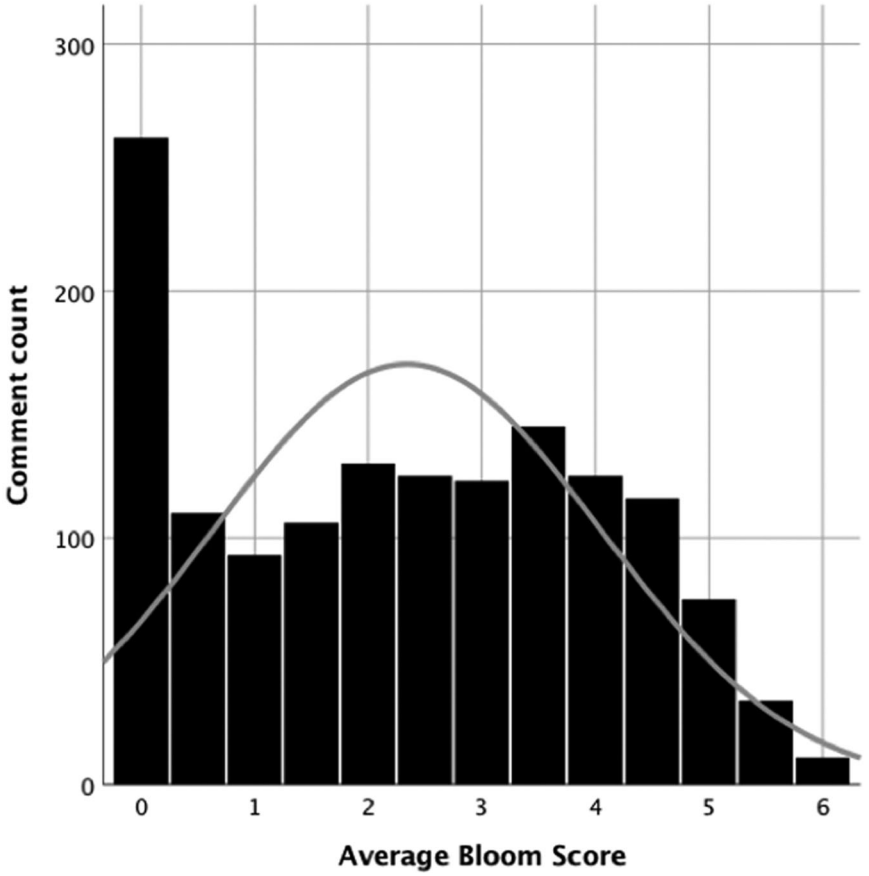

FIGURE 3 Distribution of average Bloom scores

observed, including causal words (Figure 7, because, effect, hence: $r = .573$, $p < .001$), power (control, protect, take, warn: $r = .369$, $p < .001$), and, pronouns (I, them, itself: $r = -.372$, $p < .001$).

Moderate indicators in aggregated comments only included words per sentence (wps: $r = .43$, $p < .001$), negation words (no, not, never: $r = .458$, $p < .001$), differentiation ($r = .443$, $p < .001$), cognitive process words ($r = .397$, $p < .001$), and auxiliary verbs (am, will, have: $r = .376$, $p < .001$).

Finally, word types with low correlations across all approaches to corpus analysis included conjunctions (and, also, although: $r = .28$, $p < .001$) and words with six letters or more (sixltr: $r = .2$, $p < .05$).

4.3 | Research Question 3: Social interactions as an indicator of learning

We also explored correlations between content analysis methods and measures commonly used to measure social interaction: "likes" and sentiment analysis. Although "likes" gave positive, significant results across all approaches to analysis, the correlation was weak (Table 4; maximum $r = .298$, $p < .001$). In terms of sentiment, typical measures include positive and negative emotions and emotional tone. Although all three produced significant, moderate correlations within aggregated comments, results were weakly correlated in ≥50-word comments, and negative emotion words (negemo) providing insignificant results.

5 | ANALYSIS AND DISCUSSION

In their wide-ranging review of content analysis methods de Wever et al. [15] identify six interrelated criteria for assessing their effectiveness; instruments should be "accurate, precise, objective, reliable, replicable, and valid" (p. 8). Central to these criteria is the theoretical basis of the instrument, the unit under analysis (i.e., the comment as a whole, or in part), and the extent to which they can be replicated across a variety of settings: from an individual rater agreeing with themselves, then two or more raters reaching an agreement, to the reliable use by many different groups of researchers [66]. The content analysis methods used in this study were applied by seven raters, who applied the analysis criteria to the individual, whole comments. The high level of agreement in this study suggests that these methods may be

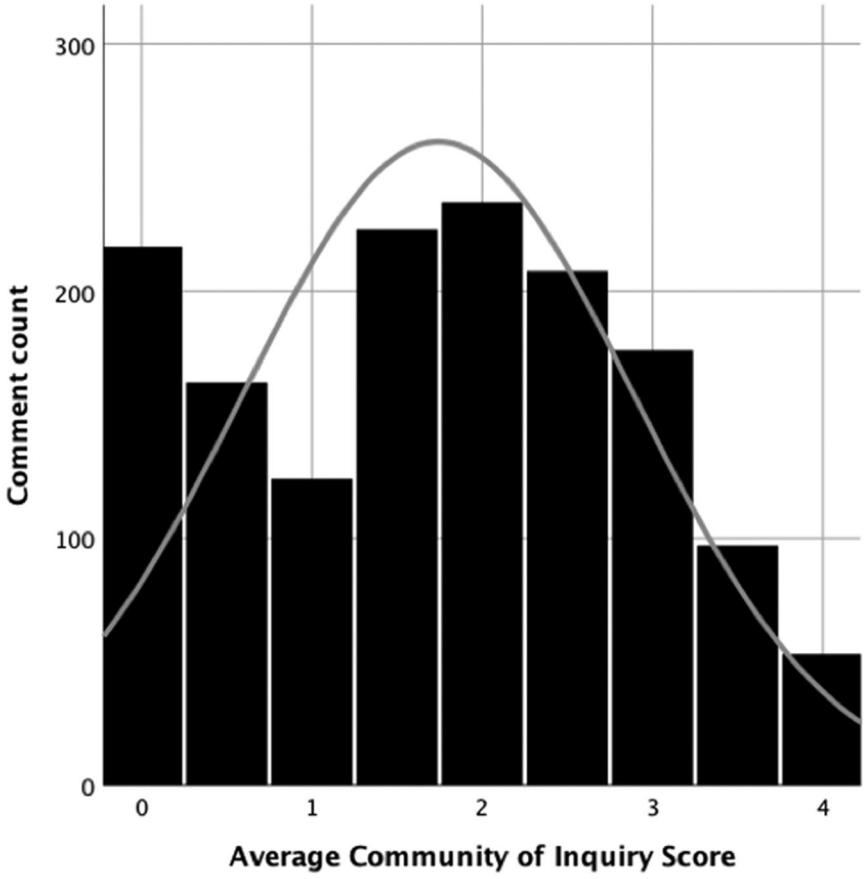

FIGURE 4 Distribution of average Community of Inquiry scores

	Number	Mean score	Standard error	Skewness	Kurtosis
Average Bloom scores	1,500	2.4	0.04	0.72	−1.17
Average CoI scores	1,500	1.7	0.03	0.25	−0.99

TABLE 3 Average Bloom and Community of Inquiry (CoI) scores distribution skewness and kurtosis statistics

successfully applied in other settings and provide the foundation of automated rating systems that closely conform to commonly held values regarding levels of critical thinking.

By exploring the comment data using three sampling techniques, we were able to investigate how the analysis methods behave in different contexts. Looking at all 1,500 comments allowed us to make inferences about general word count and interaction categories, individual ≥50-word comments facilitated LIWC word category analysis at an individual contributor level, and aggregations of all comments provided an overview of how contributors were commenting. These approaches are useful in different contexts. For example, understanding language dynamics at an individual level is important for analysing the behaviour of specific contributors, and an aggregated approach can indicate how activity within the course is generally progressing.

5.1 | Pedagogical content analysis methods

The pedagogical content analysis methods used in this study were highly correlated. As each method emphasises different aspects of cognition (Bloom—individual, CoI—distributed), this suggests that, in this study, there is a strong connection between individual levels of critical thinking and how this develops through discussion. This outcome may result from aspects of learning design that are Particular to the FutureLearn platform.

For example, though online learning environments are not always synonymous with improved critical thinking [73], there is some evidence that providing learners with timely and detailed prompts can lead to higher levels of learning [13]. By placing discussion forums within the context of each activity and providing mentor support [48], the FutureLearn platform

TABLE 4 LIWC2007 analysis

	All comments		Aggregated comments		≥50-word comments	
Category	**Bloom**	**CoI**	**Bloom**	**CoI**	**Bloom**	**CoI**
Methods	$r = .909^{***}$					
Likes	$r = .237^{***}$	$r = .243^{***}$	$r = .263^{***}$	$r = .298^{***}$	$r = .146^{***}$	$r = .149^{***}$
Positive correlations						
WC	$r = .687^{***}$	$r = .704^{***}$	$r = .759^{***}$	$r = .759^{***}$	$r = .422^{***}$	$r = .465^{***}$
Cause	$r = .125^{***}$	$r = .101^{***}$	$r = .573^{***}$	$r = .523^{***}$	$r = .224^{***}$	$r = .196^{***}$
Differentiation	$r = .220^{***}$	$r = .195^{***}$	$r = .443^{***}$	$r = .429^{***}$	$r = .100^{*}$	$r = .122^{**}$
Negation	$r = .122^{***}$	$r = .110^{***}$	$r = .458^{***}$	$r = .451^{***}$	$r = .058$ *ns*	$r = .052$ *ns*
Cogproc	$r = .125^{***}$	$r = .101^{***}$	$r = .397^{***}$	$r = .369^{***}$	$r = .164^{***}$	$r = .122^{**}$
WPS	$r = .382^{***}$	$r = .389^{***}$	$r = .430^{***}$	$r = .416^{***}$	$r = .029$ *ns*	$r = .017$ *ns*
Aux verbs	$r = .104^{***}$	$r = .092^{***}$	$r = .371^{***}$	$r = .376^{***}$	$r = .103^{*}$	$r = .052$ *ns*
Power	$r = .222^{***}$	$r = .224^{***}$	$r = .369^{***}$	$r = .358^{***}$	$r = .200^{***}$	$r = .178^{***}$
Sixltr	$r = .145^{***}$	$r = .143^{***}$	$r = .197^{*}$	$r = .200^{*}$	$r = .194^{*}$	$r = .182^{*}$
Conjunctions	$r = .271^{***}$	$r = .275^{***}$	$r = .280^{***}$	$r = .262^{***}$	$r = .100^{***}$	$r = .076$ *ns*
Negemo	$r = .112^{***}$	r = .116***	$r = .449^{***}$	$r = .445^{***}$	$r = .061$ ns	$r = .073$ *ns*
Prepositions	$r = .161^{***}$	$r = .169^{***}$	$r = .025$ *ns*	$r = .002$ *ns*	$r = .098^{***}$	$r = .115^{***}$
Negative correlations						
Pronouns	$r = -.283^{***}$	$r = -.268^{***}$	$r = -.372^{***}$	$r = -.322^{***}$	$r = -.194^{***}$	$r = -.205^{***}$
1st per. sing.	$r = -.321^{***}$	$r = -.317^{***}$	$r = -.533^{***}$	$r = -.497^{***}$	$r = -.344^{***}$	$r = -.320^{***}$
Affiliation	$r = -.119^{***}$	$r = -.117^{***}$	$r = -.387^{***}$	$r = -.331^{***}$	$r = -.095^{*}$	$r = -.127^{**}$
Posemo	$r = -.304^{***}$	$r = -.327^{***}$	$r = -.362^{***}$	$r = -.390^{***}$	$r = -.107^{**}$	$r = -.114^{**}$
Emotion	$r = -.167^{***}$	$r = -.166^{***}$	$r = -.381^{***}$	$r = -.382^{***}$	$r = -.099^{***}$	$r = -.182^{***}$

Note: Correlation coefficient (r) evaluations: Very strong–strong–moderate.

Abbreviations: CoI, Community of Inquiry; LIWC, linguistic inquiry and word count; *ns*, not significant; WC, word count; WPS, words per sentence.

*$p < .05$.

**$p < .01$.

***$p < .001$.

encourages sharing and situated debate, with the explicit intent of building communities of inquiry and inspiring higher-level learning.

In addition, the two instruments may be measuring very similar behaviours related to the depth and intensity with which people write about what they are thinking. If we agree that there is an approximate connection between the complexity of writing and depth of understanding, it is reasonable to assume that someone who has applied greater attention to their learning and wishes to share this with others, will use more elaborate arguments ("Create" in Bloom), or attempt to summarise arguments ("Resolution" in CoI), and suggests that comments evidencing these types of focus will tend to be ranked in a similar manner. Although the instruments based on those frameworks are sensitive to different aspects of learning, our results show consistency in measuring the presence and strength of critical thinking, which suggests their interchangeability in quantifying these properties in this setting.

5.2 | LIWC analysis

An important aim of this study was to determine predictors that closely align with cognitive processes in CSCL, and the literature indicates that LIWC is an accurate tool for measuring significant aspects of language use in this setting.

The most relevant outcome from regression analysis of comparisons of outputs from LIWC2015 and the content analysis instruments is the clear, statistically significant, positive correlation between word count and level of critical thinking, which confirms findings of studies which associate high word counts with thoughtful, "exploratory" exchanges in formal CSCL environments. In addition, our results for first-person person singular pronouns ("I-talk") are also supported in the literature, in showing high, significant results across all profiles, with a negative effect associated with high-level learning.

Our findings for causal, differentiation, cognitive process, and power words all provided moderate positive

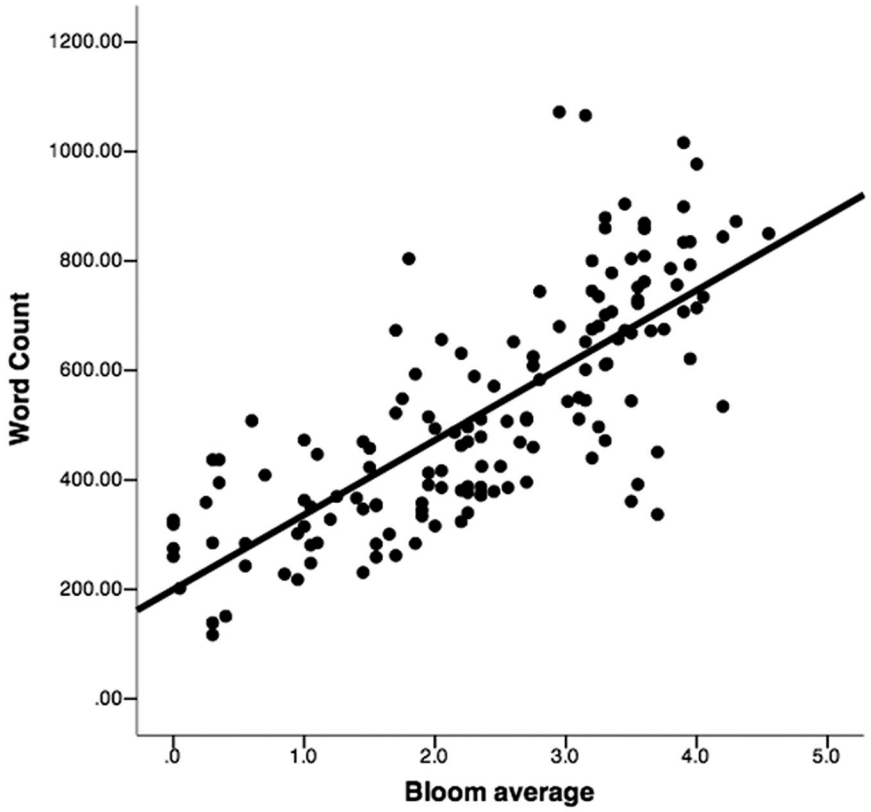

FIGURE 5 Scatter plot showing correlation between word count and Bloom score in aggregated comments ($r = .759$, $p < .001$)

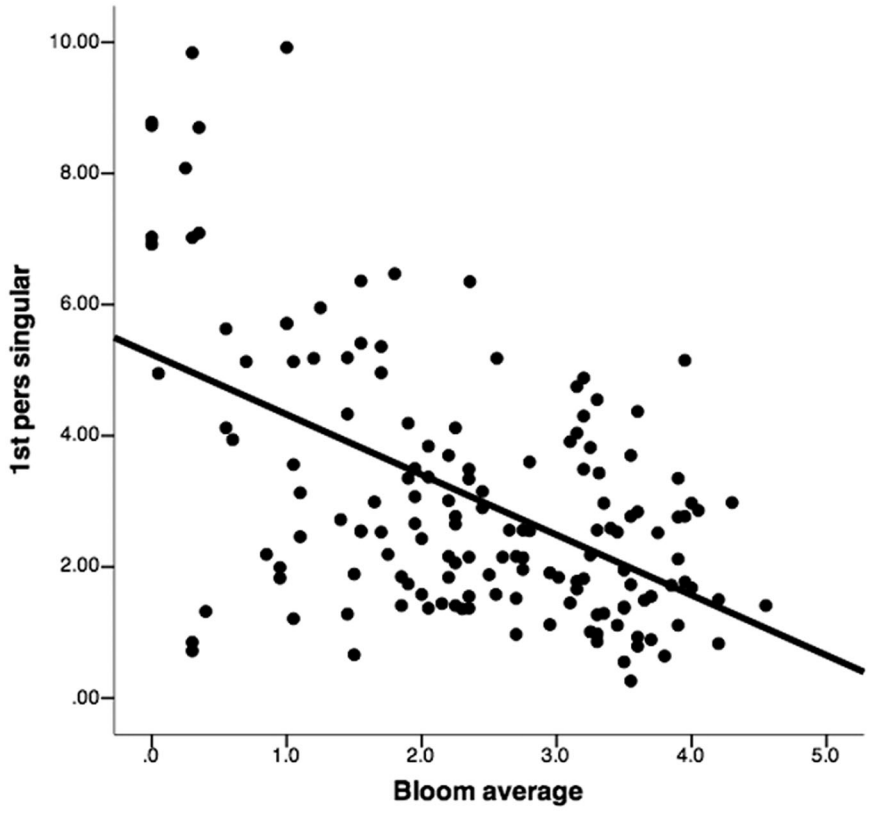

FIGURE 6 Scatter plot showing correlation between percentage of first-person singular and Bloom score in aggregated comments ($r = -.533$, $p < .001$)

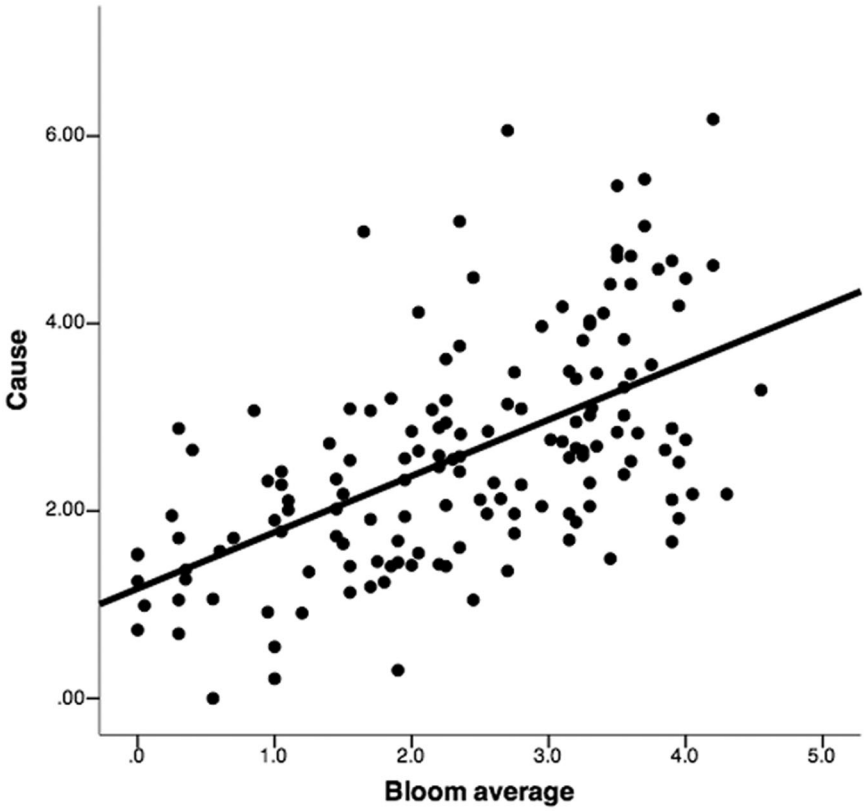

FIGURE 7 Scatter plot showing correlation between percentage of causal words and Bloom score in aggregated comments ($r = .573$, $p < .001$)

correlations with both pedagogical content analysis methods, across aggregated comments and ≥50-word comments. The findings for causal, differentiation, cognitive process accord with studies of language use in formal education as well as informal settings. The correlation of power words with higher levels of critical thinking in our study suggests self-confidence in expressing opinions.

With regard to "emotion" and positive sentiment words, the statistically significant, though moderately negative correlation between these categories and learning objects with high PS is another noteworthy outcome of this study. Some studies suggest that correlation with positive sentiment indicates loss of focus from key tasks, which, by showing a higher incidence of these categories associated with lower levels of cognitive engagement, our study appears to agree with. However, the positive association of positive sentiment words with higher levels of cognitive engagement in longer comments (containing 50 or more words) in our study, whereas weak, implies agreement with other studies that suggest that higher levels of positive language equate with a greater focus on group cohesion, and encouragement to work on-topic.

Results for negative emotion were also mixed, with significant, moderate positive correlation with the level of critical thinking in aggregated comments, but no significant results in ≥50-word comments. Although this category is associated with higher levels of critical thinking in the literature our study suggests that this, along with positive sentiment, may not be a reliable measure in all samples.

5.3 | Interaction analysis

The significant, positive association of likes in this study with high PS was unexpected, as this metric has been reported as ambiguous and unreliable, and our previous work produced insignificant results for this metric. Although results from this study, and good research practice, suggests caution when making inferences from web paradata and ambiguous phenomena like individual behaviour or cognition (especially with correlation values of less than $r = .3$), the significant result across all comments and aggregated comments implies that a process of "aggregated trustworthiness" is possibly at work [32]. In this setting, a sufficient number of MOOC forum contributors may be using the "like" button as an indicator of the trustworthiness and expertise of certain posts (rather than using it to signify agreement or ironic disagreement). Although we were unable to uncover any empirical evidence in the literature to support this particular argument, Facebook "likes" have been used to predict intelligence levels [40], and some researchers have found that the "like" feature can moderately

stimulate learner motivation [70], suggesting this may be a fruitful area for further research.

Finally, words containing six or more letters returned weak correlations, and words per sentence, negation, auxiliary verbs, conjunctions and prepositions returned mixed and insignificant results in the ≥50 word comment sample. Although the use of long words and long sentences have been associated with higher levels of critical thinking, the literature reports mixed findings for this category. In our study, negation, auxiliary verbs, and conjunctions produced moderate significant results in the aggregated comment sample, but analysis of the ≥50-word comment sample revealed no significance for these features in individual CoI coded comments, with conjunctions and auxiliary verbs producing very weak correlations in Bloom coded comments. These inconclusive results suggest that using these categories as a sole indicator of critical thinking is not advisable, but that there may be a place for these categories supporting analysis that include other significant features.

Together with the unexpected significant result for likes, the low correlation values or lack of significance of prepositions was also unanticipated. When aggregating all three MOOCs, our findings do not appear to agree with the large number of studies that have found statistically significant positive associations between prepositions and attention to reflective behaviour or increased cognitive load. However, exploratory analysis of results filtered by MOOC revealed insignificant results for prepositions in the business-related MOOC, with significant moderately correlated results for this word type in the other two. This may be explained by the very low incidence of off-topic comments in the latter MOOC samples, which further suggests that aspects of language analysis are highly context-dependent.

6 | CONCLUSIONS

This study set out to answer three key questions:

RQ1. Are pedagogical content analysis methods reliable, can different people consistently apply them, and do they identify the same types of learning activity?

Converting informal MOOC comments into comparable scores based on multiple pedagogical frameworks is a significant research challenge. In this study, a group of seven raters achieved a high degree of reliability using both pedagogical analysis methods, which enables us to have some confidence in the generalizability of these methods in future studies. Building on previous research in formal CSCL environments, we have established close associations between two distinct methods applied to informal settings, which contradicts previous findings (e.g., Reference [11]), suggesting the value of further investigation of critical thinking evaluation in MOOCs.

Although the pedagogical content analysis methods have different theoretical foundations and have been developed to assess different aspects of learning (individual and distributed cognition) when applied in this context, and correlated against language categories, sentiment and "likes", there appears to be very little difference in how they measure levels of critical thinking.

RQ2. are linguistic content analysis measures significant indicators of levels of critical thinking?

Confirming previous research (e.g., References [12,34,59]), through LIWC2015 analysis, we identified word count and first-person singular pronouns as convincing indicators of levels of critical thinking, with causal words, power and all pronouns providing moderate results. Other word categories provided mixed results within the two sampling methods used, suggesting a supporting role for these categories in future research.

RQ3. to what extent do typical measures of attention to learning indicate levels of critical thinking when applied to MOOC discussion forum comments, as identified through pedagogical content analysis?

Although producing significant results, and confirming previous work suggesting "likes" prompt engagement with higher levels of learning [13], both measures of sentiment and "likes" were weakly correlated with measures of critical thinking. Therefore, as with weakly correlated word types, this suggests secondary roles for these measures in future research.

Henri [30] suggests that the object of analysing education-based CMC interactions is to "improve the efficacy of interaction with students" (p. 117). Despite progress in codifying content analysis methods and the development of automated Natural Language Processing techniques, the absence of effective tools means the process of coding remains arduous and time-consuming [54]. For instructors, the timely identification of learners in need of pedagogical support is as relevant now as it was when Henri addressed the issue 25 years ago, and the strong correlations with LIWC2015-based proxies for pedagogical activity and the pedagogical content analysis methods in this study suggest significant promise for automated tools.

Our future work will thus involve exploring the development of real-time, automated analysis tools based

on Machine Learning techniques. Although we aim to explore a range of methods, Random Forests classifiers appear the most promising. They are among the most widely adopted classifiers in Learning Analytics research [72], offer opportunities for high accuracy and reliability [10,22] and are considered to be relatively straightforward to apply [82]. These algorithms have the potential to go beyond the linear regressions presented in this paper, combining multiple metrics to predict PS. It is our hope that such software could support learners in their personal reflection, help tutors to identify excelling or struggling students, and aid learning designers in identifying areas of weakness in their MOOCs. In the future, such tools will be essential if tutors are to effectively manage learning interactions at a massive scale.

ACKNOWLEDGEMENT

This study was supported in part by a grant from EPSRC, award: 1383089.

ORCID

Tim O'Riordan http://orcid.org/0000-0003-4905-7430

REFERENCES

1. L. K. Allen, E. L. Snow, and D. S. McNamara, *Are you reading my mind? modeling students' reading comprehension skills with natural language processing techniques*, Proceedings of the Fifth International Conference on Learning Analytics and Knowledge—LAK '15, Poughkeepsie, NY, 2015, pp. 246–254. https://doi.org/10.1145/2723576.2723617
2. L. K. Allen, E. L. Snow, and D. S. Mcnamara, *Now we're talkin': Leveraging the power of natural language processing to inform ITS development*, 7th International Conference on Educational Data Mining, 2014. London, UK, 2015, pp. 401-402.
3. C. S. C. Asterhan and M. Babichenko, *The social dimension of learning through argumentation: Effects of human presence and discourse style*, J. Educ. Psychol. **107** (2014), 740–755. https://doi.org/10.1037/edu0000014
4. C. Beleites et al., *Sample size planning for classification models*, Anal. Chim. Acta **760** (2013), 25–33.
5. J. B. Biggs, *Student Approaches to Learning and Studying*, Research Monograph, Australian Council for Educational Research, Hawthorn, Australia, 1987, p. 153.
6. B. S. Bloom, *Thought-processes in lectures and discussions*, J. Gen. Educ. **7** (1953), no. 3, 160–169. http://www.jstor.org/stable/27795429
7. B. S. Bloom, et al., Taxonomy of Educational Objectives. The Classification of Educational Goals. Handbook 1 (B. S. Bloom, ed.), McKay, New York, NY, 1956.
8. D. C. Brabham, *Moving the crowd at iStockphoto: The crowd composition of the crowd and motivations for participations in crowdsourcing application*, First Monday. **13** (2009), no. 6, 1–19. https://doi.org/10.5210/fm.v13i6.2159
9. D. W. Carroll, *Patterns of student writing in a critical thinking course: A quantitative analysis*, Assess Writ. **12** (2007), no. 3, 213–227. https://doi.org/10.1016/j.asw.2008.02.001
10. R. Caruana and A. Niculescu-Mizil, *An empirical comparison of supervised learning algorithms, ICML '06*, 23rd International Conference on Machine Learning, 2006. Pittsburgh, PA, ACM, 2006, pp. 161–168.
11. C. C. Chan et al., *Applying the Structure of the Observed Learning Outcomes (SOLO) Taxonomy on student's learning outcomes: An empirical study*, Assess. Eval. High. Educ. **27** (2002), 511–527. https://doi.org/10.1080/0260293022000020282
12. J. D. Creswell et al., *Does self-affirmation, cognitive processing, or discovery of meaning explain cancer-related health benefits of expressive writing?* Personal Soc. Psychol. Bull. **33** (2007), no. 2, 238–250. https://doi.org/10.1177/0146167206294412
13. A. Darabi et al., *Cognitive presence in asynchronous online learning: a comparison of four discussion strategies*, J. Comput. Assist. Learn. **27** (2011), no. 3, 216–227. https://doi.org/10.1111/j.1365-2729.2010.00392.x
14. A. de Liddo et al., *Discourse-centric learning analytics LAK 2011*, 1st International Conference on Learning Analytics & Knowledge, Banff, Alberta, 2011, http://eleed.campussource.de/archive/8/3336
15. B. de Wever et al., *Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review*, Comput Educ. **46** (2006), 6–28. https://doi.org/10.1016/j.compedu.2005.04.005
16. M. Delfino and S. Manca, *The expression of social presence through the use of figurative language in a web-based learning environment*, Comput. Human Behav. **23** (2007), no. 5, 2190–2211.
17. J. Dewey, *My pedagogic creed*, Sch. J. **54** (1897), no. January, 77–80. http://dewey.pragmatism.org/creed.htm
18. S. D'Mello and A. Graesser, *Dynamics of affective states during complex learning*, Learn. Instr. **22** (2012), no. 2, 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001
19. L. Dringus, *Learning analytics considered harmful*, J. Asynchronous Learn. Networks **16** (2012), no. 3, 87–100. http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ982677
20. J. D. Evans, Straightforward Statistics for the Behavioral Sciences, Brooks/Cole Publishing, Pacific Grove, CA, 1996.
21. R. Ferguson and S. Buckingham Shum, *Learning analytics to identify exploratory dialogue within synchronous text chat, LAK '11*, Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, 2011, pp. 99–103. https://doi.org/10.1145/2090116.2090130
22. M. Fernández-Delgado et al., *Do we need hundreds of classifiers to solve real-world classification problems?* J. Mach. Learn. Res. **15** (2014), 3133–3181. http://jmlr.org/papers/v15/delgado14a.html
23. U. Fischer, L. McDonnell, and J. Orasanu, *Linguistic correlates of team performance: toward a tool for monitoring team functioning during space missions*, Aviat. Space Environ. Med. **78** (2007), no. 5, Suppl., 86–95.
24. J. W. Pennebaker, M. E. Francis, and R. J. Boothe, LIWC: Linguistic Inquiry and Word Count, Erlbaum Publishers, Mahwah, NJ, 2001. https://www.researchgate.net/profile/James_Pennebaker/publication/246699633_Linguistic_inquiry_and_word_count_LIWC/links/571e3c1a08aed056fa226996/Linguistic-inquiry-and-word-count-LIWC.pdf
25. D. R. R. Garrison, T. Anderson, and W. Archer, *Critical inquiry in a text-based environment: Computer conferencing in higher*

education, Internet High. Educ. **2** (1999), 87–105. https://doi.org/10.1016/S1096-7516(00)00016-6

26. D. R. Garrison, T. Anderson, and W. Archer, *Critical thinking, cognitive presence, and computer conferencing in distance education*, Am. J. Distance Educ. **15** (2001), no. 1, 7–23. https://doi.org/10.1080/08923640109527071
27. D. Gašević, N. Mirriahi, and S. Dawson, *Analytics of the effects of video use and instruction to support reflective learning*, Proceedings of the Fourth International Conference on Learning Analytics and Knowledge—LAK '14, Indianapolis, IN, ACM, 2014, pp. 123–132. https://doi.org/10.1145/2567574.2567590
28. A. Gibson, K. Kitto, and J. Willis, *A cognitive processing framework for learning analytics*, Proceedings of the Fourth International Conference on Learning Analytics and Knowledge—LAK '14. Indianapolis, IN, 2014, pp. 212–216. https://doi.org/10.1145/2567574.2567610
29. M. Haahr and S. Haahr, Random.org, Randomness and Integrity Services Ltd., 2015, available at https://www.random.org/
30. F. Henri, *Computer conferencing and content analysis*, Collaborative Learning through Computer Conferencing. The Najadan Papers (A. R. Kaye, ed.), Springer, Berlin Heidelberg, 1992, pp. 117–136. https://doi.org/10.1007/978-3-642-77684-7_8
31. N. Indurkhya and F. J. Damerau, Handbook of Natural Language Processing (N. Indurkhya and F. J. Damerau, eds.), 2nd Edition., CRC Press, Boca Raton, FL, 2010.
32. J. Jessen and H. Jørgensen, *Aggregated trustworthiness: Redefining online credibility through social validation*, First Monday. **17** (2012), no. 1. https://doi.org/10.5210/fm.v17i1.3731
33. S. Joksimovic et al., *Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions*, Internet High. Educ. **22** (2014), 1–10. https://doi.org/10.1016/j.iheduc.2014.03.001
34. E. Kacewicz et al., *Pronoun use reflects standings in social hierarchies*, J. Lang. Soc. Psychol. **33** (2014), no. 2, 125–143. https://doi.org/10.1177/0261927X13502654
35. A. Karaksha et al., *A comparative study to evaluate the educational impact of E-learning tools on Griffith University pharmacy students' level of understanding using Bloom's and SOLO Taxonomies*, Educ. Res. Int. **2014** (2014), 1–11. http://www.hindawi.com/journals/edri/2014/934854/
36. N. Kaufmann, T. Schulze, and D. Veit, *More than fun and money. Worker motivation in crowdsourcing—A study on mechanical Turk*, Proc. Seventeenth Americas Conf. Inform. Sys., vol. 4, Detroit, MI, 2011, pp. 1–11. https://aisel.aisnet.org/amcis2011_submissions/340/
37. D. Kember, *Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow*, Int. J. Lifelong Educ. **18** (1999), no. January 2015, 18–30. https://doi.org/10.1080/026013799293928
38. M. A. Khawaja et al., *Cognitive load measurement from user's linguistic speech features for adaptive interaction design*, Lect. Notes Comput. Sci. **5726** (2009), no. LNCS (Part 1), 485–489. https://doi.org/10.1007/978-3-642-03655-2_54
39. M. A. Khawaja, F. Chen, and N. Marcus, *Using language complexity to measure cognitive load for adaptive interaction design*, Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI '10. Hong Kong, ACM, 2010, pp. 333–336. https://dl.acm.org/doi/proceedings/10.1145/1719970
40. M. Kosinski, D. Stillwell, and T. Graepel, *Private traits and attributes are predictable from digital records of human behavior*, Proc. Natl. Acad. Sci. **110** (2013), no. 15, 5802–5805. https://doi.org/10.1073/pnas.1218772110
41. V. Kovanovic et al., *What is the Source of Social Capital? The Association Between Social Network Position and Social Presence in Communities of Inquiry*, Proceedings of the Workshop Graph-Based Educational Data Mining at Educ. Data Mining Conf. (G-EDM 2014). London, UK, 2014, pp. 1–8. http://ceur-ws.org/Vol-1183/gedm_paper03.pdf
42. D. R. Krathwohl, *A revision of Bloom's Taxonomy: An overview*, Theory Pract. **41** (2002), no. 4, 212–218. https://doi.org/10.1207/s15430421tip4104_2
43. K. Krippendorff, *Reliability in content analysis: Some common misconceptions and recommendations*, Hum. Commun. Res. **30** (2004), no. 3, 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x
44. K. Krippendorff, Content Analysis: An Introduction to its Methodology, Sage, Thousand Oaks, CA/London, 2004. https://lccn.loc.gov/2003014200
45. B. Leibowitz, *Fake Facebook likes are killing social media. socialmediatoday.com blog*, 2013, available at http://www.socialmediatoday.com/content/fake-facebook-likes-are-killing-social-media
46. R. Lengelle et al., *The effects of creative, expressive, and reflective writing on career learning: An explorative study*, J. Vocat. Behav. **83** (2013), no. 3, 419–427. https://doi.org/10.1016/j.jvb.2013.06.014
47. G. Leshed et al., *Feedback for guiding reflection on teamwork practices*, Proceedings of the 2007 International ACM Conference on Supporting Group Work, GROUP '07, Sanibel Island, FL, ACM, 2007, pp. 217–220. https://dl.acm.org/doi/10.1145/1316624.1316655
48. M. León-Urrutia et al., *Mentoring at Scale: MOOC mentor interventions towards a connected learning community, EMOOCs 2015*, European MOOC Stakeholders Summit. Mons, Belgium, 2015, pp. 4. https://eprints.soton.ac.uk/373982/
49. M. Lipman, Thinking in Education, Cambridge University Press, Cambridge, UK, 2003. https://www.cambridge.org/core/books/thinking-in-education/C96667BA6F51079D8AA8D3983C57581C#
50. K. A. Meyer, *Evaluating online discussions: Four different frames of analysis*, J Asynchronous Learn. Networks **8** (2004), no. 2, 101–114.
51. J. Mezirow, *A critical theory of adult learning and education*, Experience and Learning: Reflection at Work (D. Boud and D. Walker, eds.), Deakin University, Geelong, Victoria, 1991, pp. 61–82.
52. J. Miller, *Gender, language and interaction styles in online learning environments*, 2004, available at http://www.thesisabstracts.com/ThesisAbstract_89_Gender-Language-and-Interaction-Styles-in-Online-Learning-Environments.html
53. J. L. Moore and T. R. Rocklin, *The distribution of distributed cognition: Multiple interpretations and uses*, Educ. Psychol. Rev. **10** (1998), no. 1, 97–113. https://doi.org/10.1023/a:1022862215910
54. J. Mu et al., *The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online*

discussions, Int. J. Comput. Collab. Learn. **7** (2012), no. 2, 285–305. https://doi.org/10.1007/s11412-012-9147-y

55. D. R. Newman, B. Webb, and C. Cochrane, *A content analysis method to measure critical thinking in face-to-face and computer supported group learning*, Interpers. Comput. Technol. **3** (1995), no. September 1993, 56–77. http://www.qub.ac.uk/mgt/papers/methods/contpap.html
56. S. P. Norris and R. H. Ennis, Evaluating Critical Thinking. The Practitioners' Guide to Teaching Thinking Series, Midwest Publications, Pacific Grove, CA, 1989, https://eric.ed.gov/?id=ED404836
57. C. L. Park, *Replicating the use of a cognitive presence measurement tool*, J. Interact. Online Learn. **8** (2009), no. 2, 140–155. http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ938826
58. B. F. Peden and D. W. Carroll, *Ways of writing: Linguistic analysis of self-assessment and traditional assignments*, Teach. Psychol. **35** (2008), no. 4, 313–318.
59. J. W. Pennebaker, M. Colder, and L. K. Sharp, *Accelerating the coping process*, J. Pers. Soc. Psychol. **58** (1990), no. 3, 528–537. https://doi.org/10.1037/0022-3514.58.3.528
60. J. W. Pennebaker and M. E. Francis, *Cognitive, emotional, and language processes in disclosure*, Cogn. Emot. **10** (1996), no. 6, 601–626. https://doi.org/10.1080/026999396380079
61. J. W. Pennebaker, T. J. Mayne, and M. E. Francis, *Linguistic predictors of adaptive bereavement*, J. Pers. Soc. Psychol. **72** (1997), no. 4, 863–871.
62. J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, *Psychological aspects of natural language. use: Our words, our selves*, Annu. Rev. Psychol. **54** (2003), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041
63. Pennebaker Conglomerates Inc., *LIWC2015: How it works*, 2017, available at https://liwc.wpengine.com/how-it-works/
64. J. Piaget, The Language and Thought of the Child, 3rd edn., Routledge and Kegan Paul, London, UK, 1959.
65. R. L. Robinson, R. Navea, and W. Ickes, *Predicting final course performance from students' written self-introductions: A LIWC analysis*, J. Lang. Soc. Psychol. **32** (2013), no. 4, 469–479. https://doi.org/10.1177/0261927X13476869
66. L. Rourke et al., *Methodological issues in the content analysis of computer conference transcripts*, Int. J. Artif. Intell. Educ. **12** (2001), 8–22. https://doi.org/10.1145/1518701.1518791
67. D. Rowntree, Statistics Without Tears, Penguin Books Ltd, London, UK, 1981.
68. J. B. Sexton and R. L. Helmreich, *Analyzing cockpit communications: The links between language, performance, error, and workload*, Hum. Perf. Extrem. Environ. **5** (2000), no. 1, 63–68. http://docs.lib.purdue.edu/jhpee/vol5/iss1/6
69. P. Shea et al., *Online learner self-regulation: Learning presence viewed through quantitative content- and social network analysis*, Int. Rev. Res. Open Distance. Learn. **14** (2013), 427–461.
70. R. Shih, *Can Web 2.0 technology assist college students in learning English writing? Integrating Facebook and peer assessment with blended learning*, Australas J. Educ. Technol. **27** (2011), no. Special Issue 5, 829–845.
71. G. Siemens, *Learning analytics: Envisioning a research discipline and a domain of practice*, 2nd International Conference on Learning Analytics & Knowledge (LAK 12), Vancouver, British Columbia, Canada, ACM, 2012, pp. 4–8. https://dl.acm.org/doi/10.1145/2330601.2330605
72. G. Siemens and R. S. J. D. Baker, *Learning analytics and educational data mining: towards communication and collaboration*, Proc. 2nd Int. Conf. Learn. Anal. Knowl. (2012), 252–254.
73. P. J. S. van Tryon and M. J. Bishop, *Theoretical foundations for enhancing social connectedness in online learning environments*, Distance Educ. **30** (2009), no. 3, 291–315. https://doi.org/10.1080/01587910903236312
74. D. Song, H. Lin, and Z. Yang, *Opinion mining in e-learning system*, Proceedings of the 2007 IFIP International Conference on Network and Parallel Computing Workshops, NPC 2007, Dalian, China, IEEE, 2007, pp. 788–792. https://ieeexplore.ieee.org/document/4351582
75. Y. R. Tausczik and J. W. Pennebaker, *The psychological meaning of words: LIWC and computerized text analysis methods*, J. Lang. Soc. Psychol. **29** (2010), 24–54. https://doi.org/10.1177/0261927X09351676
76. Y. R. Tausczik and J. W. Pennebaker, *Improving teamwork using real-time language feedback*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI '13, Paris, France, ACM, 2013, pp. 459–468. https://doi.org/10.1145/2470654.2470720
77. A. Timian et al., *Do patients "like" good care? Measuring hospital quality via Facebook*, Am. J. Med. Qual. **28** (2013), no. 5, 374–382. https://doi.org/10.1177/1062860612474839
78. R. Tirado, Á. Hernando, and J. I. Aguaded, *The effect of centralization and cohesion on the social construction of knowledge in discussion forums*, Interact. Learn. Environ. (2012), no. December 2012, 1–24. https://doi.org/10.1080/10494820.2012.745437
79. S. Uzuner, *Educationally valuable talk: A new concept for determining the quality of online conversations*, J. Online Learn. Teach. **3** (2007), no. 4, 400–410.
80. J. Vosecky, K. W. T. Leung, and W. Ng, *Searching for quality microblog posts: Filtering and ranking based on content analysis and implicit links*, Database Systems for Advanced Applications, **7238**, Springer, Berlin Heidelberg, Germany, 2012, pp. 397–413. https://doi.org/10.1007/978-3-642-29038-1_29
81. L. S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes, Harvard University Press, Cambridge, MA, 1978.
82. M. Walker Random Forests Algorithm. Data Science Central, 2013, available at http://www.datasciencecentral.com/profiles/blogs/random-forests-algorithm
83. Z. Waters et al., *Structure matters: Adoption of structured classification approach in the context of cognitive presence classification*, Asia Information Retrieval Societies Conference (AIRS15), Brisbane, Australia (G. Zuccon, S. Geva, H. Joho, and F. Scholer, eds.), Springer International Publishing, Cham, Switzerland, 2015, pp. 227–238.
84. R. P. Weber, Basic Content Analysis, Sage Publications, Newbury Park, CA, 1990.
85. L. Weltzer-Ward, *Content analysis coding schemes for online asynchronous discussion*, Campus-Wide Inf Syst. **28** (2011), 56–74. https://doi.org/10.1108/10650741111097296
86. M. Wen, D. Yang, and C. Rosé, *Sentiment Analysis in MOOC Discussion Forums: What does it tell us?* Proceedings of 7th International Conference on Educational Data Mining (EDM2014), London, UK, 2014, pp. 130–137. https://

educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/130_EDM-2014-Full.pdf

87. D. G. Winter et al., *Traits and motives: Toward an integration of two traditions in personality research*, Psychol. Rev. **105** (1998), no. 2, 230–250. https://doi.org/10.1037/0033-295X.105.2.230
88. A. F. Wise and T. M. Paulus, *Analyzing learning in online discussions*, The SAGE Handbook of E-Learning Research (C. Haythornthwaite, R. Andrews, J. Fransman, and E. M. Meyers, eds.), 2nd Edition., SAGE, London, UK, 2016, pp. 270–290.

AUTHOR BIOGRAPHIES

Tim O'Riordan successfully completed a PhD in Web and Internet Science at the University of Southampton, UK, in 2017. He has published in the 15th Institute of Electrical and Electronics Engineers (IEEE) International Conference on Advanced Learning Technologies and was selected for an extended presentation at the Association for Learning Technology (ALT) Annual Conference 2015. His research focuses on pedagogical, and linguistic content analysis of CSCL comment data, machine learning techniques and data visualisation.

David E. Millard is an Associate Professor of Computer and Web Science at the University of Southampton, UK, within the Web and Internet Science group in ECS in the School of Electronics and Computer Science at the University of Southampton. He is a founding member of the Web and Internet Science (WAIS) research group and is Associate Director of Research for the University of Southampton's Centre for Innovation in Technology and Education (CITE). He also is a member of the steering group for the Web Science Doctoral Training Centre (DTC) and is a member of the University of Southampton's cross-faculty Digital Economy Group. He is a Vice-chair of SIGWEB, the ACM Special Interest Group on Hypertext and the Web, and has published more than 200 papers on hypertext, the Web, and e-learning.

John Schulz is a Principal Teaching Fellow within Southampton Education School at the University of Southampton. He has published in the Higher Education Research and Development, and the Journal of Voluntary Sector Research. His research interests include organisational behaviour in nonprofit organisations and higher education institutes; and the use of technology in education. He has taught on a wide variety of programmes within the school and is currently developing an online Masters in Education.

How to cite this article: O'Riordan T, Millard DE, Schulz J. Is critical thinking happening? Testing content analysis schemes applied to MOOC discussion forums. *Comput Appl Eng Educ.* 2020;1–20. https://doi.org/10.1002/cae.22314