THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Markers of cognitive quality in student contributions to online course discussion forums

Citation for published version:
Farrow, E, Moore, J & Gasevic, D 2022, 'Markers of cognitive quality in student contributions to online course discussion forums', *Journal of Learning Analytics*, vol. 9, no. 2, pp. 38-65. https://doi.org/10.18608/jla.2022.7250

Digital Object Identifier (DOI):
10.18608/jla.2022.7250

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Learning Analytics

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 27. Oct. 2022

JOURNAL OF LEARNING ANALYTICS

Volume 0(0), 1—33. http://dx.doi.org/xxx-xxx-xxx

Markers of cognitive quality in student contributions to online course discussion forums

Elaine Farrow[1], Johanna D. Moore[2], Dragan Gašević[3]

Abstract

Through participation in asynchronous course discussion forums, students can work together to refine their ideas and construct knowledge collaboratively. Typically, some messages simply repeat or paraphrase course content, while others bring in new material, demonstrate reasoning, integrate concepts, and develop solutions. Through the messages they send, students thus display different levels of intellectual engagement with the topic and the course. We refer to this as cognitive quality. The work presented here used two widely-studied frameworks for assessing critical discourse and cognitive engagement: the ICAP and Community of Inquiry frameworks. The constructs of the frameworks were used as proxy measures for cognitive quality. Predictive classifiers were trained for both frameworks on the same data, in order to discover which attributes of the dialogue were most informative and how those attributes were correlated with framework constructs. We found that longer and more complex messages were associated with indicators of greater quality in both frameworks, and that the threaded reply structure mattered more than chronological order. By including the framework labels as additional model features, the links between frameworks were also assessed. The empirical results provide evidence that the two frameworks measure different aspects of student behaviour relating to cognitive quality.

Notes for Practice

- The Community of Inquiry and ICAP frameworks have been widely used to design and analyse student learning experiences and to understand the benefits of participation in online discussions. In previous work, the framework constructs were shown to be correlated with learning gains. We used them as independent proxy measures for the cognitive quality of student participation.
- This study looked at how various attributes of online discussions – such as text complexity and the threaded dialogue structure – were aligned with the framework constructs. We found that messages that were more deeply nested in discussion threads tended to be associated with greater quality in both frameworks. Messages that were posted later in time showed no such association. This result suggests that students should be rewarded for extending existing message threads, rather than for asking additional novel, but unrelated, questions.
- We also found that the frameworks were not closely aligned with each other, suggesting that they measure different aspects of student experience in online discussions. Thus, using their constructs in combination in future studies would be expected to provide richer insights, compared to using either one alone.

Keywords

discussion forum, participation, engagement, quality, Community of Inquiry, cognitive presence, ICAP, feature analysis

Submitted: 05/06/20 — **Accepted:** 18/03/22 — **Published:** xx/xx/xx

Corresponding author [1] *School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB UK. Email: Elaine.Farrow@ed.ac.uk ORCID ID: 0000-0002-1152-3443*

[2] *School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB UK. Email: J.Moore@ed.ac.uk ORCID ID: 0000-0001-7247-6823*

[3] *Faculty of Information Technology, Monash University, Clayton 3800, Australia. Email: Dragan.Gasevic@monash.edu ORCID ID: 0000-0001-9265-1908*

1. Introduction

Discussion forums are widely used across all types of learning environments, from traditional face-to-face classroom settings to distance learning and MOOCs (Garrison, 2011, 2016; Wise et al., 2016). The messages exchanged in discussion forums

constitute a particularly valuable source of information about student learning. Forums allow students to engage socially as well as intellectually, and provide scope for problem solving and discussion (Garrison et al., 2000). Students can engage with one another and with instructors. They can also write simply to clarify their own thoughts. Earlier work showed that learning can take place through the process of articulating knowledge and engaging in reasoning (Ferguson et al., 2013), including self-explanation (Chi et al., 1989). In the practical inquiry model (Garrison et al., 2001), both the 'private world' of reflection and the 'shared world' of discussion play a vital role in learning. Work on MOOCs (Wang et al., 2015, 2016; Wise & Cui, 2018) showed that participation in the optional discussion forums was positively correlated with learning gains, even though messages often received only a nominal reply.

It is increasingly common that the number of messages generated in a discussion forum is too large for instructors to monitor effectively. Many studies have used content analysis, both manual and automatic, to label discussion forum messages according to multi-level theoretical frameworks for measuring critical thinking and cognitive engagement (McKlin, 2004; Corich et al., 2006; Waters et al., 2015; Kovanović et al., 2014; Gašević et al., 2015; Wang et al., 2015, 2016; Kovanović et al., 2016; R. Ferreira et al., 2018; Neto et al., 2018; Farrow et al., 2019; Yogev et al., 2018; Taskin et al., 2019; M. Ferreira et al., 2020; Hu et al., 2020). If student contributions to the discussion can be assessed automatically while the course is still in progress, this could allow instructors to identify students who are bored, frustrated, or struggling, or lessons which cause confusion, while there is still time to intervene. For example, discussion forum transcripts could be colour-coded to indicate how the conversation was progressing, enabling instructors to see at a glance where to direct their attention (Yogev et al., 2018).

Our aim in this study was to identify attributes of the dialogue that could be used in an automated system to discriminate between contributions of varying quality, as measured by both the *phase of cognitive presence*, defined in the Community of Inquiry framework (Garrison et al., 2000), and the *mode of cognitive engagement*, defined by the ICAP framework (Chi & Wylie, 2014). The use cases targeted by this work are primarily practical in outlook. Rather than aiming to construct and test a novel process model of discourse, we instead chose to use two well-established, validated, frameworks to provide independent proxy measures of the quality of student participation in the discussion activity. By discovering empirically which attributes of the dialogue are linked to higher quality contributions in these two frameworks, we can offer guidance on how to design better discussion forums. This is important because some of the measures captured by dialogue attributes are commonly specified in the protocols for discussion forum participation, such as the number of questions to be asked and answered, and the frequency of posting (Gilbert, 2002). It is instructive to see how those measures are correlated with the framework constructs – and thus, by implication, with message quality – in practice. Our findings could guide educators to develop participation requirements that are better aligned with the desired learning outcomes, by fostering deeper engagement with the intellectual content of the discussion. The dialogue attributes that are identified as being most indicative of quality might also inform the hints and suggestions in future automated feedback systems, where full content analysis is impractical.

This study builds on earlier work that successfully automated the labelling process for the constructs of the Community of Inquiry framework and identified classification features that were most relevant for distinguishing between them (Kovanović et al., 2016; Neto et al., 2018; M. Ferreira et al., 2020). Our choice of a second framework to use in the present study was guided by several considerations. Most importantly, we wanted to maintain the focus on the students' *cognitive* engagement with the content of the course. Other recent studies have looked at automatic labelling of help-seeking behaviour in forums (Cross et al., 2017) and taxonomies of questions asked by students (Harrak et al., 2018), but these were too different from our primary focus. Studies identifying content-related posts and threads (Cui & Wise, 2015; Wise et al., 2016) were also relevant, but we were looking for more than a binary flag. None of these alternative frameworks accounted for the central role played by the *interactions* between the students in a discussion forum that allow them to construct knowledge collaboratively. The ICAP framework met all our requirements. Its main focus is cognitive, it has been widely used in other educational settings, and previous studies have shown that automated labelling can be applied successfully.

We first treated each of the frameworks separately, then went on to consider whether the gold-standard labels from one framework might be useful as features in a predictive model trained to assign labels from the other framework – thereby revealing overlaps and contrasts between the framework constructs. This work is an extension of a previous study looking at the link between dialogue attributes and measures of participation (Farrow et al., 2020). In that earlier work we looked at each of the frameworks independently. In this paper, we also considered how constructs from one framework might inform the assignment of labels from the other framework. We looked at the importance of those constructs relative to the other dialogue attributes in the model. This approach allowed us to observe similarities and differences between the frameworks and to identify potential alignments between individual CoI phases of cognitive presence and ICAP modes of cognitive engagement. Knowing whether, and how, the frameworks align could allow researchers to reinterpret the results of previous studies using the combined theoretical insights from the two frameworks; and could improve the development of future automated classifiers by expanding and refining the set of indicators for any aligned constructs.

We contribute to the existing body of literature on assessing student participation in online discussions in two ways: 1) We identify dialogue attributes that inform the cognitive quality of student contributions, according to the constructs of two

widely-used theoretical frameworks. 2) We provide empirical evidence that these two frameworks measure different aspects of participation, and thus, that educators and researchers would gain a richer understanding of student behaviour by combining insights from both of them. This is the first work we know of that has used both of these popular theoretical frameworks together.

2. Literature Review

In this section, we introduce the two frameworks used in this study (Sections 2.1 and 2.2) and review a selection of relevant previous studies where automated methods were used to label the framework constructs based on attributes of the dialogue. We consider similarities and differences between the frameworks (Section 2.3) in terms of the conceptual approach, granularity, and purpose of each of them. We conclude the section by presenting and motivating our two research questions (Section 2.4).

As we set out in Section 1, our intention in this work was to use the two frameworks as independent quality measures. We acknowledge that the frameworks were developed to address different learning situations, and that discussion forums themselves can be used for different purposes within a course, including reflection, debate, and problem-solving. Our particular interest is the type of setting where a discussion forum allows students to exercise critical thinking and build knowledge collaboratively (Garrison et al., 2001; Corich et al., 2007). In such a setting, it is desirable that students should engage deeply with the intellectual content of the subject matter, integrate content from other sources, and share their thoughts and discoveries with their peers. However, it is not straightforward to quantify the extent to which students are achieving this aim. We posit that the constructs of the two frameworks we have chosen, described in detail below, can provide useful proxy measures for the cognitive quality (or simply *quality*) of student contributions.

2.1 The Community of Inquiry framework

The Community of Inquiry (CoI) framework for online education is a powerful tool for analysing and developing effective learning experiences (Garrison et al., 2000; Garrison, 2016). The framework identifies three main elements ('presences') that are important for a successful educational experience: a social environment conducive to learning (*social presence*), a well-designed course with ongoing facilitation (*teaching presence*), and the student's own cognitive engagement with the subject matter (*cognitive presence*).

In this work, we focus specifically on cognitive presence as a measure of the quality of student participation. Of the three presences, cognitive presence is considered to be the most fundamental to learning. It has four phases:

- **Triggering Event:** the initial question that sparks a discussion.
- **Exploration:** the phase of the discussion when many new ideas are being considered.
- **Integration:** the phase where ideas begin to coalesce into a more coherent form as connections are identified.
- **Resolution:** the final phase, where a conclusion has been reached, perhaps in the form of a hypothesis that can be tested.

It is desirable for a discussion to progress through all four phases of cognitive presence (Figure 1), although not every discussion will do so. There is an expectation that progression through the phases will be somewhat ordered in time, because the phases build on one another. The initial *Triggering Event* sets the context, and messages in the *Exploration* phase will tend to address it directly. The *Integration* phase might bring together several points from the *Exploration* phase. If the *Resolution* phase is reached, relevant messages may refer back to *Exploration* messages and even to the *Triggering Event*. Additional *Triggering Event* messages, perhaps in the form of clarification questions, can form part of a message thread and can open up new lines of relevant discussion. A single long message might also contain content belonging to several phases, demonstrating one student's own progression of ideas. The coding scheme indicates that these should be coded with the highest phase (Garrison et al., 2001). This is referred to as *coding up*. Messages with no sign of cognitive presence are coded as *Other*.

The CoI framework was designed as a "practical approach to judging the nature and quality of critical discourse in a computer conference" (Garrison et al., 2001, p. 7). It is a general purpose framework, applicable to multiple domains. Although CoI was originally developed with traditional credit-bearing courses in mind, a revised rubric for manual labelling of messages with the CoI phases of cognitive presence was recently applied to messages from a MOOC discussion forum (Hu et al., 2020). Agreement between the two annotators was high (Cohen's $\kappa = 0.93$). In the updated scheme, a message that paraphrased information previously given was labelled as a *Triggering Event*, as were messages that affirmed or disagreed with a previous point but without giving any reasons. Agreements and disagreements that were underpinned by reasoning were labelled as *Integration*.

CoI has been widely used to analyse student learning in online courses, and predictive models have been developed for identifying its elements automatically using the text of discussion forum messages (McKlin, 2004; Corich et al., 2006;

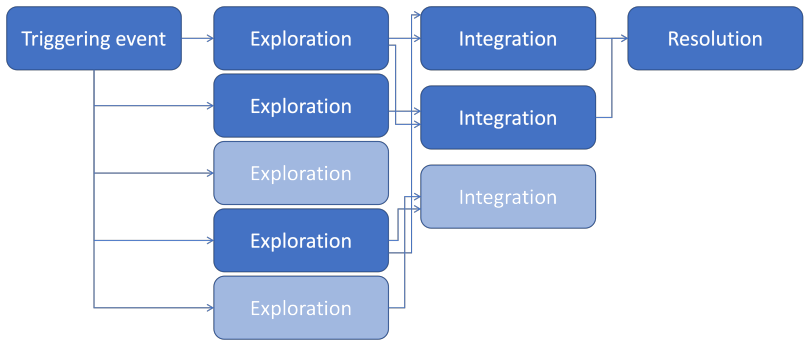


Figure 1. An idealised example of how a discussion might progress through the CoI phases of cognitive presence.

Kovanović et al., 2014; Waters et al., 2015; Kovanović et al., 2016; Neto et al., 2018; Farrow et al., 2019; Hu et al., 2020; Barbosa et al., 2020). Gašević et al. (2015) investigated the effect of varying the scaffolding provided to different cohorts of students, an aspect of teaching presence. Recent work by M. Ferreira et al. (2020) developed models for labelling the indicators of social presence automatically. However, the bulk of prior work in this area has focused specifically on cognitive presence.

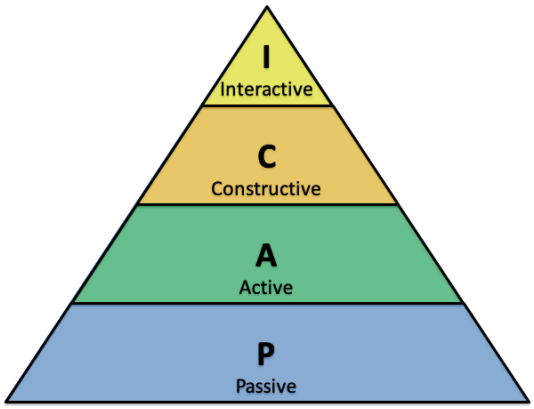
Early work (McKlin, 2004) used neural networks to detect the phases of cognitive presence automatically through content analysis (Cohen's $\kappa = 0.70$). Inputs were mainly dictionary-based features, along with features describing the position of the message in the threaded discussion. Corich et al. (2006) labelled the sentences within messages, rather than whole messages, using an automated content analysis tool. Kovanović et al. (2014) used support vector machines to label messages with CoI phases of cognitive presence. They used standard bag-of-words text features without adding any features to account for context. Waters et al. (2015) incorporated context by using conditional random fields to generate a sequence of labels for each message thread, rather than considering messages in isolation. Messages with multiple replies were thus analysed repeatedly as part of several sequences, and labelled using majority vote.

More recently, several groups of researchers used random forests (Breiman, 2001) to model the distribution of discussion forum messages in a corpus (Kovanović et al., 2016; Neto et al., 2018; Yogev et al., 2018; Farrow et al., 2019; Barbosa et al., 2020; M. Ferreira et al., 2020). Yogev et al. (2018) found random forests to be superior to other methods including logistic regression and support vector machines. The random forest approach has the benefit of being a 'white-box' method that allows inspection of its workings. Analysis of the most predictive model features can provide further insight into factors affecting cognitive presence. A study by Kovanović et al. (2016) used a random forest together with both structural and high-level linguistic features. It claimed high accuracy (Cohen's $\kappa = 0.63$), but a later replication study (Farrow et al., 2019) indicated that the model's predictive power was likely to have been over-estimated (Cohen's $\kappa = 0.38$). As a consequence, it may have seen some features as more predictive than was really the case, while disregarding others that actually have more discriminative power. A similar approach was applied to discussion forum messages written in Portuguese (Neto et al., 2018), achieving Cohen's $\kappa = 0.72$. Recent work (Barbosa et al., 2020) trained a random forest using English language data and used it to label data in Portuguese, based on dialogue attributes that were available for both languages, and reporting Cohen's $\kappa = 0.53$.

2.2 The ICAP framework

The ICAP framework (Chi & Wylie, 2014) takes a different approach from CoI, defining cognitive engagement based on overt, observable, behaviours. The framework looks at how learning activities relate to students' cognitive engagement with the learning materials. It can be applied to in-person activities as well as to those conducted online. Like CoI, it is not domain-specific. Four modes of engagement are identified, and the framework predicts that higher modes will be correlated with greater learning gains. The four modes, in descending order, are **I**nteractive, **C**onstructive, **A**ctive, and **P**assive. Each of these modes represents a qualitatively different *kind* of growth in knowledge, not simply a bigger or smaller change. Nevertheless, each mode subsumes the modes below it (Figure 2). Off-task behaviours do not constitute any sort of cognitive engagement.

Unlike the CoI phases of cognitive presence, there is no expectation of a temporal progression through the ICAP modes of cognitive engagement during a discussion or other learning activity. In fact, the modes are most commonly seen as relating to the learning activity as a whole, rather than to individual student actions. In some cases, participants in the same activity might demonstrate different modes of cognitive engagement. For example, in a large group discussion, those who speak and build on each other's contributions are using the *Interactive* mode, while others who make summary notes may be using the *Constructive* mode, or perhaps the *Active* mode if their notes are verbatim, and those who only listen demonstrate *Passive* engagement. As with the *coding up* approach described in Section 2.1, the assigned label relates to the highest mode of engagement that is observed. An important point to note is that an activity is only considered to be *Interactive* if the student in question is responding directly to another student's contribution. If the student is responding to another source, such as a textbook or a



- **Interactive:** extending or challenging the *Constructive* ideas of a partner.
- **Constructive:** generating novel output that relates to course content, beyond what was given.
- **Active:** engaging in some activity that is related to the course and requires focused attention.
- **Passive:** reading or watching course materials without actively doing anything else.

Figure 2. The hierarchical modes of cognitive engagement in the ICAP framework.

pre-recorded video, then the activity will be labelled as *Constructive*.

Despite the original focus on a learning activity taken as a whole, prior work has demonstrated the feasibility of applying a modified version of the ICAP schema to label the individual messages from MOOC discussion forums (Wang et al., 2015, 2016) and to student comments on an annotated electronic course text (Yogev et al., 2018) and on MOOC videos (Taskin et al., 2019). Atapattu et al. (2019) completely automated the initial labelling process for student contributions in a MOOC.

In the first study using the ICAP framework to identify higher-order thinking behaviours in MOOC discussion forum data (Wang et al., 2015), messages were labelled using categories corresponding to the *Active*, *Constructive*, and *Interactive* modes. Linear regression was used to identify significant correlations between students' post-test results and flags indicating whether any of the messages posted by a particular student were labelled as *Active*, *Constructive*, or *Interactive*. Results showed that learning gains were significantly greater for students who posted *Active* and *Constructive* messages. Posting an *Interactive* message was only significant in cases where the student had posted fewer than 3 messages (the median) in total. The results thus did not support the original ICAP hypothesis but instead found that *Active* behaviours were the strongest predictors of learning gains and *Interactive* messages the weakest. However, any combination of flags could be present for each student – no attempt was made to *code up*.

The coding scheme was revised in later work (Wang et al., 2016) and the updated coding manual was offered as a research output. The label definitions were revised with reference to the original ICAP framework and adapted for MOOC forum data. An extended label set was introduced that allows for finer-grained distinctions between messages within two of the modes: *Constructive* mode was divided into *Constructive Reasoning* and *Constructive Extending*, while *Active* mode was divided into *Active Targeted* and *Active General*.

- **Constructive Reasoning:** elaborating on a point or displaying reasoning about course content.
- **Constructive Extending:** proposing ideas, sharing resources, or asking questions going beyond course materials.
- **Active Targeted:** referring explicitly to course content by paraphrasing or asking clarification questions.
- **Active General:** showing other signs of being actively engaged with the course.

The finer-grained modes were not used directly for analysis but instead recombined into *Constructive* and *Active*, respectively. Genuinely *Interactive* messages were very rare. Most messages were self-contained and did not evoke contentful responses from others. In order to investigate the relationship between the framework constructs and learning outcomes, students who posted *Interactive* and/or *Constructive* messages were grouped together and contrasted with those who only posted *Active* messages and with those whose messages were all *Off-task*. Being in the *Off-task* group was significantly associated with lower post-test results. Membership of the group with *Interactive* and/ or *Constructive* messages had a larger effect size than membership of the *Active* group.

The ICAP framework has also been used to analyse student comments added to an annotated electronic course textbook (Yogev et al., 2018) – a setting quite different from a typical discussion forum, because the content under discussion can be highlighted and annotated directly. The interface allowed students from courses in physics and biology to add their comments and questions directly alongside the source material. These annotations were labelled using an adapted version of the coding manual that was developed in Wang et al. (2016), using the message as the unit of analysis. The label definitions were updated to take into account the context provided by the highlighted section of the course text as well as the other students' annotations.

Treatment of messages of affirmation (that is, agreement or thanks) varied depending on their context. In many cases, an affirmation message inherited the label from the message to which it was responding. During the analysis, the labels were assigned numeric values and average values computed. This contrasts with the *coding up* approach discussed earlier. Students who wrote more annotations than others mainly generated questions and thus ended up with a lower average score.

Manual labelling of discussion forum posts with framework constructs can be laborious and time consuming. One approach to automating the labelling used word embeddings (Le & Mikolov, 2014) to measure semantic similarity in order to distinguish between *Constructive* and *Active* contributions in the context of a community-centric MOOC for teachers' professional development (Atapattu et al., 2019). In that study, the unit of analysis was the student rather than the message. Participants whose contributions were highly similar to the course materials were labelled as *Active*, while those that had little overlap with the course materials were labelled *Constructive*, on the assumption that they were generating new knowledge. Manual content analysis was used to validate a selection of the contributions that had been automatically tagged as *Constructive*, with similarity scores in the lowest quartile. Of the 67 examples considered, only one lacked any *Constructive* contributions. No attempt was made to investigate whether other *Constructive* contributions were missed in those contributions labelled *Active*.

Video-based content plays an important role in many online learning environments. One video-based learning platform was enhanced with 'nudges' to prompt students to engage more by writing comments attached to specific portions of the video (Taskin et al., 2019). This was the first study to apply the ICAP framework to interactions with video material. Although more comments were written in the 'nudge' condition, these were typically shallow and simple (*Active* mode) and did not improve learning gains. There was no correlation between the quantity of *Constructive* comments and higher learning gains, but students who wrote at least one *Constructive* comment had a higher gain than those who did not write any. The number of *Constructive* comments written was negatively correlated with students' extrinsic motivation scores, indicating that students with high extrinsic motivation would tend to respond to the nudges by writing simpler *Active* comments. Messages of affirmation were assigned to the *Active* category. There was no scope for interaction in this video-based activity.

The distribution of ICAP modes of cognitive engagement varied between the learning activities reviewed in this section based on the task characteristics. For example, in the studies using MOOC discussion forums (Wang et al., 2015, 2016), very few posts were labelled as *Interactive*; while in the video annotation study (Taskin et al., 2019) they were entirely absent. This variation is both expected and welcome, since it reinforces previous findings (Chi & Wylie, 2014) that the ICAP framework is applicable to a broad range of learning situations.

2.3 Similarities and differences between the frameworks

The two frameworks used in this study have several obvious similarities. Both consider the cognitive aspects of student behaviour to be the most important for learning. Both emphasise the value of building on the contributions of other participants in the discussion, through integration of ideas. In practical terms, automated approaches that assign framework labels to new data have been developed for both frameworks and are generally applied at the granularity level of individual messages.

However, while both frameworks address student learning, they do so from different perspectives. They were developed independently from each other and with different goals in mind. CoI was developed specifically in order to understand the benefit of computer-mediated education and to explain how students develop their ideas through discussion leading to social knowledge construction (Garrison et al., 2000; Garrison, 2011). ICAP has a broader scope and has been demonstrated to be effective in predicting the educational value of several different interventions, in a classroom setting as well as online (Chi & Wylie, 2014). There are other obvious differences. The ICAP framework was originally used to predict the educational effectiveness of a particular type of learning activity (Chi & Wylie, 2014), while the CoI framework describes the contributions of students during different phases of a computer-mediated discussion (Garrison et al., 2001). In terms of framework structure, ICAP has a single set of modes, while CoI includes two additional presences, social and teaching, that support the development of cognitive presence. Another notable difference is that the CoI phases of cognitive presence are expected to develop in order as the discussion progresses, with each phase building on the content of previous phases, from *Triggering Event* to *Resolution*; whereas the ICAP modes of cognitive engagement do not have any in-built ordering. That is, regardless of what has gone before, a student may legitimately respond with a message that is labelled *Active*, *Constructive*, or *Interactive* at any time.

Previous work comparing CoI with Bloom's Taxonomy (Bloom et al., 1956) and the SOLO taxonomy (Biggs & Collis, 1982) found that the higher-level labels were often correlated across all three frameworks, while the lower-level labels tended to be more diverse (Schrire, 2006). A study looking at doctoral-level classes in educational leadership (Meyer, 2004) used CoI and Bloom's Taxonomy and found a similar proportion of messages at the highest levels in both frameworks. Finding the commonalities and differences in how the CoI and ICAP frameworks apply to one specific data set offers a useful contribution to the theoretical understanding of online learning and learning through discussion. If the two frameworks are found to measure broadly the same things, then results derived using each of them in previous studies could also be expected to be applicable to work using the other. For example, learning interventions that encouraged discussion participants to progress to the higher phases of cognitive presence (*Integration* and *Resolution*) would be expected to demonstrate greater learning gains in the same

way as interventions that targeted the higher ICAP modes of cognitive engagement. In our recent work (Farrow et al., 2021), we used a network analytic approach to quantify the associations between the CoI and ICAP frameworks and measure the moderation effects of two instructional interventions on those associations.

It might be helpful to consider what an optimally structured discussion thread would look like in each of the frameworks. In terms of the CoI phases of cognitive presence, the messages nearer the top level of the idealised thread would explore the topic widely (*Exploration*). The mid-level messages would bring together different ideas and give reasons for selecting some and rejecting others (*Integration*). Messages describing potential conclusions or solutions (*Resolution*) would be nested most deeply. In the best case, an individual student would contribute at all of the levels, not just championing their own ideas but also showing how they relate to others. In terms of the ICAP modes of cognitive engagement, the idealised thread would feature interactions between multiple students (not just a back-and-forth); each message would address specific points from previous contributors, mainly using reasoning to support or contradict them, and sometimes combining points from more than one source; and many of the messages would incorporate supporting facts from outside sources. In the best case, most of an individual student's contributions would be *Interactive*, with few *Constructive* messages that ignore the contributions from others and very few *Active* messages that add information without giving reasons for its relevance.

Taking both frameworks together, in the optimal case we would expect to see sub-threads started by multiple students, proposing different ways to address the topic (*Exploration* + *Constructive*). Follow-on messages in each sub-thread would support or contradict those proposals using reasoning and evidence (*Exploration* + *Interactive*). Some would bring together ideas from multiple sub-threads (*Integration* + *Interactive*). Ideally, this would eventually lead to a developing consensus around one or more possible solutions (*Resolution* + *Interactive*). There might be some brief social messages interspersed along the way, providing encouragement and appreciation (*Other* + *Affirmation*).

2.4 Research questions

We addressed two research questions in this study. First, we considered each framework on its own and investigated the relationship between dialogue attributes and framework labels. It is reasonable to suppose that some attributes of a dialogue might vary systematically between messages with different framework labels. Investigating these relationships would give insight into factors that relate to the quality of participation. Our first research question was therefore,

- **RQ1:** *What is the relationship between the dialogue attributes and the framework labels for the CoI phases of cognitive presence and the ICAP modes of cognitive engagement?*

Our second research question looked at the potential explanatory power of labels from one framework in relation to the other, as a means of investigating potential links between the frameworks. If a label from one framework was in fact a close analogue of a label in the other, we would expect to find that each one was highly predictive of the other. If instead the concept represented by a single label from one framework corresponded to multiple labels in the other framework, we would expect to see improvements in the performance of a predictive model that assigns framework labels to new data when the constructs from the second framework were added as model features. For example, if the behaviour captured by the *Interactive* mode of ICAP were distributed across the *Exploration*, *Integration*, and *Resolution* phases of cognitive presence, we would expect to see that those phases were highly predictive features in a model predicting the ICAP modes of cognitive engagement. Conversely, if adding those features did not lead to an improvement in model performance, that would indicate that the frameworks are measuring different aspects of the learning experience. In that case, labelling data with both sets of labels in future studies would give a richer picture of student participation. In practical terms, such an approach is increasingly feasible as automated methods of labelling are developed.

Our expectation was that, while there would be some similarities, overall the frameworks would provide complementary views on the learning experience. Specifically, we expected to find that the higher-level labels were more closely aligned than the lower ones, as in prior work that compared frameworks that can be used for assessment of quality in online discussions (Schrire, 2006). Building constructively on the contributions of others is the defining feature of *Interactive* mode, and some messages in both the *Integration* and *Resolution* phases could be expected to meet that criterion. The definition of the *Exploration* phase, with its focus on bringing in new ideas, seems conceptually most closely related to the novelty expected in *Constructive* mode. Some messages in the *Integration* phase may also be expected to fit here. Messages in *Active* mode refer to existing content without adding anything substantially new; these would perhaps relate most strongly to the *Triggering Event* phase, where previous statements are questioned or paraphrased (Hu et al., 2020).

Comparing the explanatory power of framework constructs alongside the dialogue attributes indicates not only the relationships between constructs, but also their relative strength and importance. Thus, our second research question asked,

- **RQ2:** *What is the explanatory value of the labels from one framework when modelling the other, in the context of the dialogue attributes examined in RQ1?*

3. Methodology

In order to address both of our research questions, we used a data set of course discussion forum messages that was annotated with labels relating to the constructs from the two frameworks we are examining. We describe the data set in Section 3.1, the framework labels in Section 3.2, and the dialogue attributes in Section 3.3. In common with previous work, the unit of analysis was the message.

Previous work (Kovanović et al., 2016; Neto et al., 2018; Barbosa et al., 2020) showed how random forest models can be used to compare dialogue attributes by examining their predictive value when they are used as model features. The relative importance of different classification features can be discovered by inspecting the Mean Decrease Gini (MDG) for each feature (Breiman, 2001). MDG is calculated as the mean of the decrease in the Gini impurity measure across all decision tree nodes where the feature is used. The MDG score is designed to estimate the importance of each feature independently. If two features are closely correlated, their scores would be expected to be similar. When a random forest is constructed, only a subset of the available features is used for each of the trees in the forest. Therefore, there will be some trees where the most highly predictive features are not used, and this allows us to observe the behaviour of the other features in their absence. In this way, random forests overcome one of the main shortcomings of decision trees: in the presence of highly correlated features, a single decision tree will make an arbitrary choice about which to use. In contrast, if two features in a random forest model are highly correlated with one another, both will be included in the model and will receive similar MDG scores.

Our first experiment (Section 3.4) addressed RQ1 by training several multi-class random forest models using cross-validation and using the best of these models to assign labels to the messages in a held-out test set. Having determined that the predictive performance of the models was sufficiently good, we examined which of the dialogue attributes that were used as model features could best discriminate between messages in terms of the five labels from the CoI framework (four phases of cognitive presence plus *Other*) and the four modes of engagement in the ICAP framework.

In order to answer RQ2, our second experiment (Section 3.5) looked at whether using the gold-standard labels from one framework as additional model features improved the ability of a model to correctly assign the labels from the other framework. We ranked the relative importance of the features in order to determine how the framework constructs compared with the dialogue features used in Experiment 1.

3.1 Description of the data

This work makes use of a data set that has previously been used in several studies of cognitive presence (Kovanović et al., 2014; Gašević et al., 2015; Waters et al., 2015; Kovanović et al., 2016; Farrow et al., 2019). It was collected from a fully online distance-learning course at a Canadian university that formed part of a Masters degree in software engineering. We use data from six course offerings, which took place between 2008 and 2011. The messages were exchanged during the first of four graded assignments, described below. The distribution of participants and messages across the sessions is shown in Table 1.

Table 1. Statistics for the 6 course offerings used in this work.

Session	Participants		Messages	
Winter 2008	16		212	
Fall 2008	24		633	
Spring 2009	12		243	
Fall 2009	9		63	
Winter 2010	15		359	
Winter 2011	13		237	
Average (SD)	14.8	(5.1)	291.2	(192.4)
Median (Q1–Q3)	14.0	(12.3–15.8)	240.0	(218.3–330.0)
Total	89		1747	

Each student created and shared a video presentation based on a research paper relevant to the course, then started a new thread in the discussion forum to host a conversation around their presentation. Every student was expected to lead one discussion thread and contribute to at least three others. We do not have access to the presentations themselves, only to the text-based discussion that followed. Students were in general highly motivated, since the first assignment accounted for 10% of the final course mark. In contrast to the MOOC data used in Wang et al. (2016) and Taskin et al. (2019), genuine interaction between students was commonplace in this course.

3.2 Labels assigned by the frameworks

The messages in our data set had previously been annotated with their phase of cognitive presence by two expert coders (98.1% agreement, Cohen's $\kappa = 0.974$). Table 2 shows the distribution of the CoI phases of cognitive presence across the data.

Table 2. Breakdown of messages by CoI phases of cognitive presence. The *Other* class was used for messages that displayed no sign of cognitive presence.

Cognitive presence phase	Label	Example behaviour	Count	Percentage
Triggering Event	**triggering**	Recognising an issue or asking a question that sparks discussion	308	17.63%
Exploration	**exploration**	Exploring and exchanging information, considering new ideas	684	39.15%
Integration	**integration**	Identifying connections between ideas and constructing meaning	508	29.08%
Resolution	**resolution**	Testing a hypothesis or reaching a consensus	107	6.12%
Other	**other**	Commenting without any signs of cognitive presence	140	8.01%
All			1747	100.00%

For this study, we additionally annotated each message with a label indicating the relevant mode of cognitive engagement from the ICAP framework. In the labelling task itself, we built on earlier work (Wang et al., 2015, 2016; Yogev et al., 2018) that developed guidelines for labelling ICAP constructs in data from MOOC discussions and annotated course texts using an extended label set. Like Yogev et al. (2018), we adapted the label definitions to take account of context beyond the student messages themselves. For them it was the textbook material and in our case it was the video presentations. Messages that referenced content from the presentation were treated differently from those that referenced an earlier message: only those in the second case could be labelled *Interactive*.

The original ICAP hypothesis predicts that learning experiences employing the higher modes will lead to deeper understanding (Chi & Wylie, 2014). The work that introduced the extended label set (Wang et al., 2016) did not present any theoretical claims or predictions about the new finer-grained labels. The order of precedence in the annotation guidelines implies that *Constructive Reasoning* is considered to be higher than *Constructive Extending*, and similarly that *Active Targeted* is higher than *Active General*. This ordering was later explicitly stated by Yogev et al. (2018). Neither study used the finer-grained modes directly for analysis, but instead recombined them into *Constructive* and *Active*.

Messages of affirmation, consisting primarily of agreement or thanks expressed in response to an earlier message, have been treated in a variety of ways in earlier studies. Taskin et al. (2019) considered them to indicate 'shallow' engagement. Hu et al. (2020) distinguished between agreements which were justified with reasoning, and those which stood alone; the former were considered to demonstrate deeper cognitive presence than the latter. Yogev et al. (2018) treated simple messages of affirmation – without any explanation or reasoning – as a special case: the label assigned to them depended on the label of the earlier message to which they were responding. If the earlier message was labelled as *Interactive* or *Constructive Reasoning*, then the affirmation message was labelled as *Constructive Extending*; in all other cases, the affirmation message simply inherited the earlier label. However, for the purpose of developing an automated classifier that can label future data reliably, it is preferable to assign each label based only on attributes of the current message. Otherwise, two affirmation messages with identical content (*e.g.*, "Thanks for your reply") and appearing in the same position within a thread could receive different labels depending on the labels of the earlier messages. Therefore, in the current work, we did not assign the derived label to affirmation messages directly. Instead, we gave them the *Affirmation* label as a placeholder. Once all the messages in the data set have had labels assigned (by manual coding or using an automated classifier), a simple rule-based transformation can be applied to relabel all *Affirmation* messages, based on the labels that were assigned to the messages they are affirming.

Each message in our data set was assigned a single label, corresponding to the highest mode of cognitive engagement that was identified in the message, similar to the *coding up* process that was used for the CoI labels. For this purpose, the *Affirmation* label was considered to sit between *Constructive Extending* and *Active Targeted* (Yogev et al., 2018). The full extended ICAP label set used for annotation, along with the distribution of the labels across the data set, is presented in Table 3. We had no access to data indicating when a student read a message without responding, so the *Passive* label was not used.

Two postgraduate students worked independently as annotators to label the data for this study. One was closely involved in the research and labelled the full data set while revising and refining the annotation guidelines to address specific features of this data set. The second annotator was provided with the annotation guidelines and received initial training. Both annotators then iteratively labelled the messages from 5 discussion threads and resolved disagreements through discussion, following which the annotation guidelines were clarified further.[1] A further 12 discussion threads (202 messages) were labelled independently and the labels from these were used to assess inter-annotator agreement as Cohen's $\kappa = 0.623$, indicating 'substantial' agreement (Landis & Koch, 1977). Overall, 20.1% of the discussion threads and 17.1% of the messages in the corpus were labelled independently by two people. Messages were always considered for labelling one thread at a time, to allow annotators to distinguish between novel contributions and paraphrases of earlier messages (Ferguson et al., 2013).

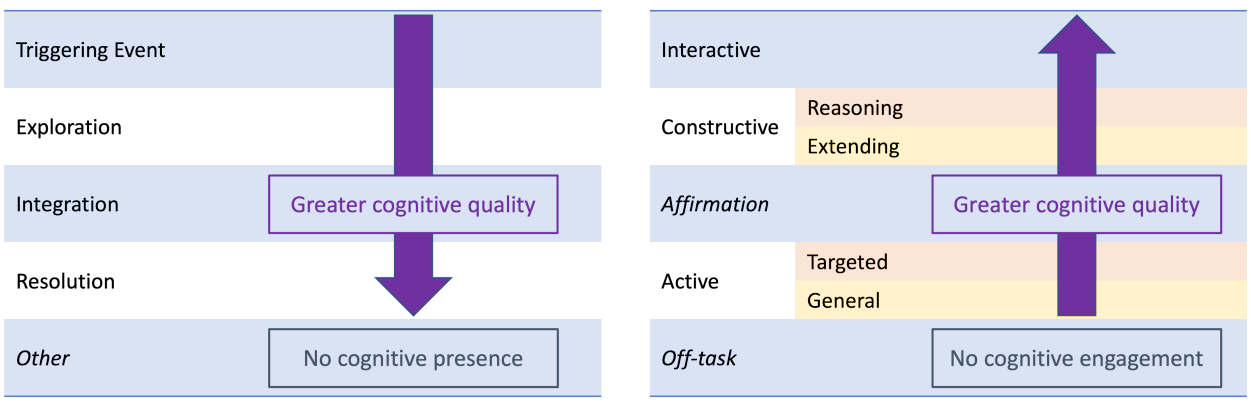
The two sets of framework constructs are shown together in Figure 3. In subsequent Figures, they are shown with quality

[1] The revised annotation guidelines are available from `https://homepages.inf.ed.ac.uk/efarrow/#Resources`.

Table 3. Breakdown of messages across modes in the extended cognitive engagement taxonomy, adapted from Yogev et al. (2018) and based on the ICAP framework. The *Affirmation* class is not part of the framework and was added by the authors of the current paper, as described in the text. The *Off-task* class contains messages displaying no cognitive engagement.

Cognitive engagement mode	Label	Example behaviour	Count	Percentage
Interactive	**I**	Displaying explanation or reasoning about the current topic in response to an earlier message	579	33.14%
Constructive Reasoning	**C1**	Displaying explanation or reasoning about the current topic	313	17.92%
Constructive Extending	**C2**	Introducing new content to the discussion	409	23.41%
Affirmation	**F**	Affirming what was said in an earlier message	73	4.18%
Active Targeted	**A1**	Referencing specific previous content	75	4.29%
Active General	**A2**	Showing other signs of being engaged with course content	287	16.43%
Passive	**P**	Reading messages without responding	0	0.00%
Off-task	**O**	Commenting without any relation to the current topic or the course	11	0.63%
All			1747	100.00%

increasing in the same direction, for ease of comparison. Figure 4 shows a heat map indicating the distribution of labels across the two frameworks. Redacted examples of messages with each framework label are included in Table 20 in the Appendix.



(a) The CoI phases of cognitive presence

(b) The extended ICAP modes of cognitive engagement

Figure 3. The framework constructs used to label the data, with an indicator showing the direction corresponding to greater cognitive quality. Note that the fine-grained subcategories of the *Constructive* and *Active* modes were not addressed by the original ICAP hypothesis; their relative positions within the hierarchy are therefore tentative, as is that of the *Affirmation* label.

3.3 Dialogue attributes used as model features

Each message was automatically annotated with 206 classification features. A total of 91 word counts were derived using the LIWC software package (Tausczik & Pennebaker, 2010), along with 106 metrics related to text coherence, complexity, readability, and lexical category use from Coh-Metrix (McNamara et al., 2014). These 197 lexical features are the same as those used in earlier work using this data (Kovanović et al., 2016; Farrow et al., 2019). In addition, we defined 9 structural features capturing aspects of the discussion structure, described next and shown in Table 4.

The threaded nature of the forum means that every message can receive multiple replies, and replies can themselves receive replies. A new reply can be added at any level in the chain at any time. Without knowing which messages a student has actually read, we need to make some assumptions. A message posted as a reply to another message can be expected to relate to that message in a meaningful way. Similarly, the impact of a message on the discussion can be measured not only by the number of replies it gets, but perhaps also by the total count of replies-to-replies: that is, counting all the descendant messages. We thus defined three features related to the position of the message in the thread (depth in thread, first message, and last message) and two features for replies (number of direct replies and total number of replies).

We expected that the chronological order of messages would also be relevant, so we ordered the messages within each thread using time-stamp order and then derived features using that ordering (position from start, position from end, and fractional position). A final feature (discussion size) captures the total number of messages in the thread, allowing the classifier to distinguish between longer and shorter discussions. While the value of this feature would be the same for all messages that

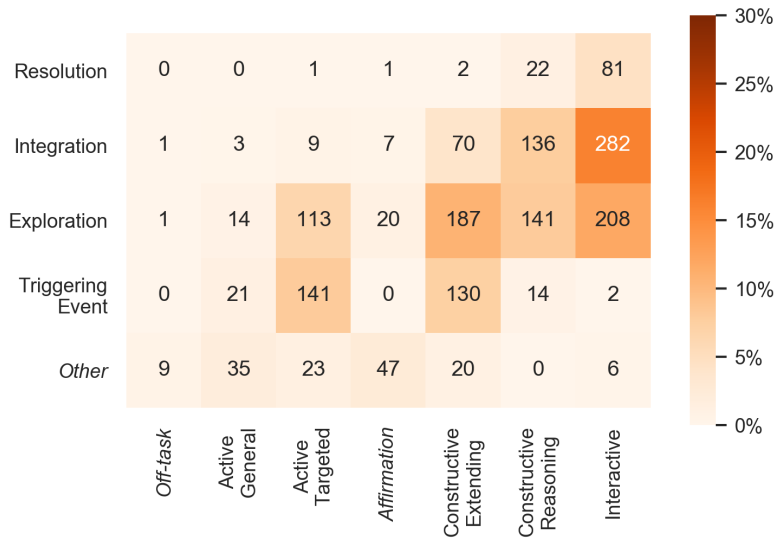


Figure 4. Cross-tabulation of labels across the two frameworks. The numbers in each cell indicate the raw number of messages having that combination of framework labels and the colour density indicates the distribution of labels within the data set.

Table 4. Structural attributes derived from the data and used as model features.

Structural attribute	Feature	Description and Rationale
Depth in thread	`message.depth`	The depth of the message within the threaded view of the discussion. A deeper message is more likely to be an example of *Interactive* mode.
First/ last message	`message.is.first`, `message.is.last`	Binary indicators for the first and last message in a discussion thread, defined chronologically. By definition, the first message in a thread cannot be *Interactive*.
Number of direct replies	`message.replies.direct`	The number of direct replies to the message. Messages relating to *Triggering Events* and *Exploration* are expected to generate more replies than those in deeper phases (Waters et al., 2015).
Total number of replies	`message.replies.all`	The cumulative number of direct and indirect replies (replies to replies). Exchanges with a partner are a key feature of the *Interactive* mode. This feature may also capture the role of *Triggering Events* and *Exploration* better than direct replies alone (McKlin, 2004).
Position from start	`message.pos.start`	The index of the message in chronological order from the beginning of the discussion. Early messages may be more likely to introduce new material.
Position from end	`message.pos.end`	The index of the message in chronological order from the end of the discussion. Later messages may build on what has gone before to achieve greater quality.
Fractional position	`message.pos.frac`	The position of the message chronologically within the discussion, as a fraction of the total discussion size. This feature seeks to allow for natural variations in discussion length.
Discussion size	`message.thread.size`	The total number of messages in the current discussion. We hypothesise that a short discussion is less likely to progress to deeper phases of cognitive presence than a longer one.

were part of the same thread, it could prove informative at a later point in the decision tree, perhaps after the messages had already been divided using their position within the threaded or chronological order.

3.4 Experiment 1

Experiment 1 addressed RQ1. We used the first five offerings of the course as training data for a multi-class random forest classifier, and kept back the data from the final session as unseen test data with which to assess the best model. This is in line with best-practice on replicability (Gardner et al., 2018), since a course changes every time it runs and a useful model needs to be general enough to make predictions on future runs of the course.

For the first experiment, we recombined the finer-grained distinctions within the *Constructive* and *Active* modes, in common with prior work (Wang et al., 2016; Yogev et al., 2018). As there were so few *Off-task* messages, those records were excluded from our analysis of the ICAP framework. The breakdown of CoI phases of cognitive presence within the training and test partitions is shown in Table 5, while the figures for the ICAP modes of cognitive engagement are shown in Table 6.

Table 5. Messages by CoI phases of cognitive presence in training and test partitions.

	Training		Test	
Cognitive presence phase	Count	Percentage	Count	Percentage
Triggering Event	280	18.54%	28	11.81%
Exploration	608	40.26%	76	32.07%
Integration	425	28.15%	83	35.02%
Resolution	85	5.63%	22	9.28%
Other	112	7.42%	28	11.81%
All	1510	100.00%	237	100.00%

Table 6. Messages by ICAP modes of cognitive engagement in training and test partitions. The messages labelled as *Off-task* were excluded from the analysis.

	Training		Test	
Cognitive engagement mode	Count	Percentage	Count	Percentage
Active	313	20.73%	47	19.83%
Affirmation	66	4.37%	9	3.80%
Constructive	616	40.79%	106	44.73%
Interactive	506	33.51%	73	30.80%
Off-task	*9*	*0.60%*	*2*	*0.84%*
All	1510	100.00%	237	100.00%

We explored 20 different settings for the `mtry` parameter that controls how many of the 206 classification features were available as candidates at each decision tree split point. The specific values to be tested were automatically determined by the `caret` library in `R` based on the number of features in the model; here, they were 2, 12, 23, 34, 44, 55, 66, 77, 87, 98, 109, 120, 130, 141, 152, 163, 173, 184, 195, and 206. For each `mtry` setting, we trained 1,000 trees and used 10-fold cross-validation, repeated 10 times, to select the best performing value. A final random forest model was built using this value and data from the full training set.

The number of data points belonging to each outcome class (*i.e.* the phases of cognitive presence and the ICAP modes) was unbalanced (Tables 5 and 6). It is well-known that unbalanced data can cause problems for classification techniques. Rebalancing the classes in the training data can alleviate these problems, but it is not guaranteed to improve the performance of the model. For this reason, we also compared models trained directly on the unbalanced training data against models using SMOTE (Synthetic Minority Over-sampling TEchnique (Chawla et al., 2002)) to rebalance the classes in the outcome variable such that every outcome class had the same size (Table 7). Following best practice, the SMOTE algorithm was run inside the cross-validation loop (Farrow et al., 2019). With 10-fold cross-validation, 90% of the data is used for training at each step and the remaining 10% is used for validation. SMOTE was applied to the training data for each fold; new synthetic data points were generated from the smaller classes until every class had the same number of examples. By augmenting only the training data and leaving the test data unchanged, we avoided a source of data contamination that could lead to the selection of a sub-optimal model (Farrow et al., 2019). A summary of the models trained in Experiment 1 is shown in Table 7.

For each framework, the model that achieved the highest Cohen's κ score in cross-validation was used to assign labels to the held-out test data from the final run of the course, and the relative importance of each variable in the model was compared.

Table 7. Models trained in Experiment 1 for each of the two frameworks, with and without class rebalancing.

Model	Outcome classes	Model features			
		LIWC	Coh-Metrix	Structural	Total count
CoI phases of cognitive presence					
Model CoI-1a	5 phases, original distribution	91	106	9	206
Model CoI-1b	5 phases, rebalanced with SMOTE	91	106	9	206
ICAP modes of cognitive engagement					
Model ICAP-1a	4 modes, original distribution	91	106	9	206
Model ICAP-1b	4 modes, rebalanced with SMOTE	91	106	9	206

In this way, we identified the dialogue attributes that were best able to distinguish between the different CoI phases of cognitive presence and ICAP modes of cognitive engagement.

3.5 Experiment 2

In our second experiment, in order to address RQ2, we used the constructs from each of the frameworks as additional classification features when training a model to label new data using the other framework. This experiment mimics the situation where a data set has already been labelled using one framework and we wish to add the other set of labels automatically. It allows us to discover the relative explanatory value of the framework constructs compared with the existing dialogue attributes and to see how labels from the two frameworks are aligned, as discussed in Section 2.4.

We trained a model to label the CoI phases of cognitive presence (five labels including *Other*) using the gold-standard ICAP label for each message as a model feature alongside the same dialogue attributes that were used in Experiment 1 (Model CoI-2a). We were interested to discover how useful the ICAP constructs were for improving the predictive performance of the model, and how important they were relative to the dialogue attributes. The *Off-task* label was included as a feature in this analysis, despite its rarity. Model CoI-2b again labelled the 5 cognitive presence classes, but used as features the extended set of ICAP labels shown in Table 3, excluding only the unused *Passive* label. This allowed us to discover whether the additional finer-grained detail improved the model's predictive power.

We also trained two models to predict the ICAP modes of cognitive engagement using the dialogue attributes used in Experiment 1 with the gold-standard CoI phases of cognitive presence label as an additional feature. Here, the *Off-task* messages were once again excluded from the outcome variable, since it would be unreasonable to expect a model to learn to identify a class adequately from so few examples. Model ICAP-4 was trained to predict the 4 ICAP modes, while Model ICAP-6 used those same labels to predict the extended set of 6 ICAP modes. These models and their features are summarised in Table 8.

The labels identifying framework constructs were not used directly as model features. Instead, each one was expanded to create a collection of binary features, one for each possible value of the label – a 'one-hot' encoding. For example, in Model CoI-2a and Model CoI-2b, for each message that had the label **I**, indicating *Interactive* mode, the feature `ICAP-I` was assigned the value 1, while the features corresponding to the other ICAP labels were all assigned the value 0. Similarly for Model ICAP-4 and Model ICAP-6, if a message had the label **exploration**, the binary `CoI-exploration` feature would have the value 1, while the features corresponding to the other CoI labels would all have the value 0.

Table 8. Models trained in Experiment 2. The features used were the same 206 dialogue attributes as in Experiment 1, plus framework constructs expanded into a one-hot representation.

Model	Outcome classes	Model features			
		Dialogue attributes	ICAP modes	CoI CP phases	Total count
CoI phases of cognitive presence					
Model CoI-2a	5 phases, rebalanced with SMOTE	206	5	—	211
Model CoI-2b	5 phases, rebalanced with SMOTE	206	7	—	213
ICAP modes of cognitive engagement					
Model ICAP-4	4 modes, rebalanced with SMOTE	206	—	5	211
Model ICAP-6	6 extended modes, rebalanced with SMOTE	206	—	5	211

We prepared the data as before, using the first five offerings of the course as training data and using the final session as unseen test data with which to assess the predictive power of the best model. We used the SMOTE algorithm inside the cross-validation loop to rebalance the outcome variable classes such that every outcome class had the same size. We explored

20 settings for the `mtry` parameter in the same way as in Experiment 1. For each `mtry` setting, we trained 1,000 trees and used 10-fold cross-validation, repeated 10 times, to select the best performing value. For each framework, we chose the model that performed best in cross-validation and trained a final random forest model using the best `mtry` value and data from the full training set. This was used to assign labels to the held-out test data. The relative importance of each variable in the model was compared, in order to discover the relative importance of the framework constructs compared to the dialogue attributes.

4. Results

4.1 Experiment 1

In Experiment 1, we addressed RQ1 and investigated the relationship between the dialogue attributes and the framework labels.

4.1.1 Predictive performance metrics

When dealing with unbalanced classes, as we were in this study, Cohen's κ and the macro-averaged F_1 score are more informative than accuracy. We chose the best model for each framework based on Cohen's κ (Table 9). In each case, we found that rebalancing the outcome classes using SMOTE inside the cross-validation loop gave better results during training than using the original unbalanced data, in common with prior work (Farrow et al., 2019).

Table 9. Cross-validation results for Experiment 1: outcome metrics and the best value for the `mtry` tuning parameter, with and without class rebalancing using SMOTE. The best results for each framework are in bold.

Model	Outcome classes	Best `mtry`	Macro F_1	Cohen's κ
CoI phases of cognitive presence				
Model CoI-1a	5 phases, original distribution	77	0.500	0.416
Model CoI-1b	5 phases, rebalanced with SMOTE	55	**0.542**	**0.429**
ICAP modes of cognitive engagement				
Model ICAP-1a	4 modes, original distribution	109	0.684	0.619
Model ICAP-1b	4 modes, rebalanced with SMOTE	34	**0.687**	**0.624**

We used the best model from each framework to assign labels to the held-out data from the final offering of the course. The Cohen's κ scores from the test data are shown in Table 10, along with the Precision, Recall, and F_1 scores for each class of the outcome variable. For the model that labels the CoI phases of cognitive presence, a Cohen's κ of 0.358 indicates a 'fair' level of agreement with the gold-standard human coding, while the Cohen's κ of 0.694 for the ICAP modes of cognitive engagement indicates 'substantial' agreement (Landis & Koch, 1977). The macro-averaged F_1 score for the CoI model was 0.515, and for the ICAP model it was 0.764. The macro-averaged F_1 score for Model CoI-1b was particularly affected by the low F_1 score for the *Resolution* phase, for which there were few training examples (5.63% of the training data). In contrast, the F_1 score for the *Affirmation* mode in Model ICAP-1b was only slightly lower than for the other modes, despite it accounting for only 4.37% of the training data.

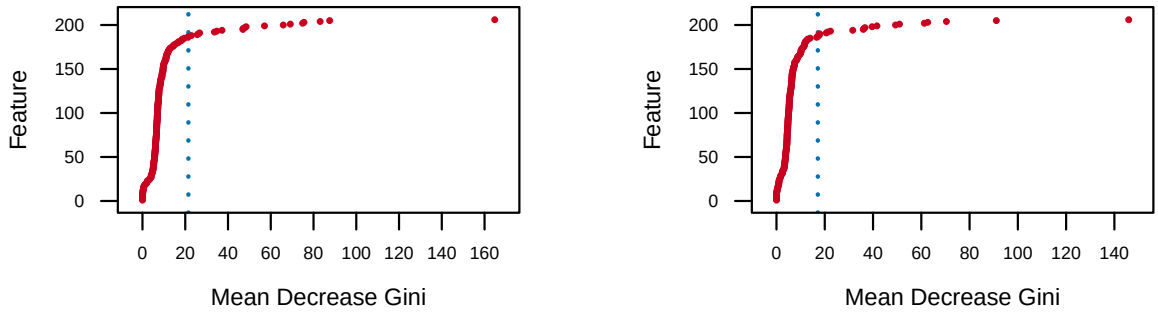
Table 10. Experiment 1: Outcome metrics on the held-out test data.

Outcome classes	Precision	Recall	F_1	Macro F_1	Cohen's κ
CoI phases of cognitive presence, using Model CoI-1b				0.515	0.358
Triggering Event	0.719	0.821	0.767		
Exploration	0.460	0.526	0.491		
Integration	0.543	0.530	0.537		
Resolution	0.250	0.136	0.176		
Other	0.640	0.571	0.604		
ICAP modes of cognitive engagement, using Model ICAP-1b				0.764	0.694
Active	0.762	0.681	0.719		
Affirmation	0.571	0.889	0.696		
Constructive	0.857	0.793	0.824		
Interactive	0.778	0.863	0.818		

4.1.2 Analysis of variable importance

The best models showed that, in each case, a small subset of features had a high degree of explanatory power, evidenced by their high Mean Decrease Gini (MDG) values (Figure 5). The top 20 features by importance for each framework are listed in

Tables 11 and 12. Their distributions across the CoI phases of cognitive presence and the ICAP modes of cognitive engagement are plotted in Figures 7, 8, and 9 in the Appendix, where the features are listed in alphabetical order for ease of comparison.



(a) Model CoI-1b **(b)** Model ICAP-1b

Figure 5. Mean Decrease Gini indicating variable importance in the best models for each framework in Experiment 1. In both cases, the vertical dotted line separates the top 20 features.

Considering first the **lexical features** that appear in both lists, we see that longer messages and fewer question marks were associated with higher cognitive quality in both frameworks. The number of words in the message (`cm.DESWC`) appears in the top 5 for both models. Similarly, lower levels of lexical diversity, measured by type-token ratio (`cm.LDTTRa` and `cm.LDTTRc`), were associated with deeper phases of cognitive presence and also with deeper cognitive engagement. In contrast, when using the alternative VOCD lexical diversity metric that aims to compare texts of different lengths more reliably (`cm.LDVOCD`), the relationship was reversed: *higher* levels of lexical diversity were seen to be associated with both deeper cognitive presence and cognitive engagement. These results are in line with prior work on CoI (Kovanović et al., 2016; Neto et al., 2018). It is interesting to see that they apply to ICAP as well. The number of expressions of positive emotion (`liwc.posemo`) and the number of affective process words (`liwc.affect`) also appear in the top 15 in both lists. Both are strongly indicative of *Other* messages (those that display no signs of cognitive presence) and of *Affirmation* messages.

Looking at the **structural features** that appear in both lists, messages that were more deeply nested in the discussion – although not necessarily posted later in time – were more likely to come from the *Resolution* or *Other* phases of cognitive presence, and to indicate the *Interactive* or *Affirmation* mode of ICAP (Table 13). None of the features relating to time-stamp order within a thread appeared in the top 20 for ICAP, and the only one to do so for CoI was the 'first message' indicator, which is strongly indicative of *Triggering Event*s (McKlin et al., 2001). However, the number of replies (both direct and indirect) a message received was highly predictive for both frameworks. The values were highest for messages labelled *Constructive* and *Triggering Event* respectively. Both of these observations can be explained by noting that the way the original discussion task was structured meant that the first message in each thread was nearly always assigned the same label: *Triggering Event* or *Constructive* (specifically, *Constructive Extending*).

Moving on to features that are predictive for one of the two frameworks but not the other, we see that messages displaying deeper levels of the CoI phases of cognitive presence used more words from the LIWC categories relating to discrepancies (`liwc.discrep`, words such as *should* and *would*) and money (`liwc.money`, words such as *owe*). The length of the discussion thread (`message.thread.size`) appeared at position 14 in Table 11, but the distribution of its values did not vary systematically across the phases. Meanwhile, the Coh-Metrix measure tracking the amount of 'given' versus 'new' information in each sentence within a message (`cm.LSAGN`) was highly predictive for the ICAP modes of cognitive engagement. The highest values were seen for *Interactive* messages, which are expected to build on and develop the arguments from earlier messages; and lowest for *Affirmation* messages. The number of expressions of assent (`liwc.assent`) was, unsurprisingly, highest on average for the *Affirmation* mode. We also note that the use of personal pronouns (`liwc.ppron`) is indicative of *Active* mode, where quoting is expected, and of *Affirmation* messages.

Overall, the most predictive lexical features have similar MDG scores compared with the most predictive structural features, in both Table 11 and Table 12. Finally, we observe that the features appearing at positions 10 and 11 in Table 12 (`cm.DESPL` and `cm.DESSC`) had identical distributions across the outcome classes. In fact, since every message in our data set was formatted as a single paragraph of text, they were actually measuring the same thing: the mean length of a paragraph in sentences, and the total number of sentences in the message. The small discrepancy in their MDG values was simply due to the random nature of the random forest. The fact that both of these highly correlated attributes were found to be equally predictive reinforces one of the benefits of random forests, discussed in Section 3 – no arbitrary choice is made.

Table 11. The 20 most important features in Model CoI-1b, the best CoI cognitive presence model for Experiment 1, ranked from most to least important by Mean Decrease Gini (MDG). Features that also appear in the top 20 for the best ICAP model (Table 12) are shown with a coloured background.

Rank	Feature	Description	MDG
1	`cm.DESWC`	Number of words	164.76
2	`message.is.first`	First message	87.54
3	`liwc.posemo`	Number of +ve emotion words	83.11
4	`cm.WRDMEAc`	Meaningfulness	75.55
5	`message.depth`	Message depth in discussion	74.90
6	`cm.LDTTRa`	Lexical diversity, all words	69.15
7	`liwc.SemiC`	Number of semi-colons	65.84
8	`liwc.QMark`	Number of question marks	57.07
9	`cm.WRDHYPn`	Hypernyms for nouns	48.46
10	`message.replies.direct`	Number of direct replies	47.84
11	`liwc.affect`	Number of affective process words	47.05
12	`liwc.discrep`	Number of discrepancy words	46.98
13	`liwc.money`	Number of money words	37.18
14	`message.thread.size`	Discussion size	34.78
15	`message.replies.all`	Total number of replies	33.64
16	`cm.LSASSpd`	SD of LSA overlap in paragraph	26.71
17	`cm.DESWLltd`	SD of word length in letters	26.10
18	`cm.LDVOCD`	Lexical diversity, VOCD	25.64
19	`cm.LDTTRc`	Lexical diversity, content words	22.89
20	`liwc.hear`	Number of hearing-related words	21.47

Table 12. The 20 most important features in Model ICAP-1b, the best ICAP model for Experiment 1, ranked from most to least important by Mean Decrease Gini (MDG). Features that also appear in the top 20 for the best ICAP model (Table 11) are shown with a coloured background.

Rank	Feature	Description	MDG
1	`message.depth`	Message depth in discussion	145.94
2	`liwc.assent`	Number of expressions of assent	91.08
3	`message.replies.direct`	Number of direct replies	70.39
4	`message.replies.all`	Total number of replies	62.65
5	`cm.DESWC`	Number of words	61.24
6	`cm.LDVOCD`	Lexical diversity, VOCD	50.97
7	`cm.LDTTRa`	Lexical diversity, all words	49.32
8	`liwc.posemo`	Number of +ve emotion words	41.61
9	`cm.LSAGN`	LSA given-new ratio	39.59
10	`cm.DESPL`	Mean length of paragraphs	36.66
11	`cm.DESSC`	Number of sentences	36.21
12	`liwc.QMark`	Number of question marks	35.95
13	`liwc.affect`	Number of affective process words	31.59
14	`cm.LDTTRc`	Lexical diversity, content words	22.39
15	`liwc.Period`	Number of periods	21.22
16	`cm.RDFKGL`	Flesch-Kincaid Grade Level score	20.60
17	`liwc.tentat`	Number of tentative words	17.88
18	`cm.RDFRE`	Flesch Reading Ease score	17.86
19	`cm.DESWLltd`	SD of word length in letters	17.42
20	`liwc.ppron`	Number of personal pronouns	17.13

Table 13. Mean and standard deviation of selected message position attributes for each of the CoI phases of cognitive presence and ICAP modes of cognitive engagement in Experiment 1.

	Depth in thread `message.depth`		Position from start `message.pos.start`		Fractional position `message.pos.frac`	
Outcome classes	mean	(SD)	mean	(SD)	mean	(SD)
CoI phases of cognitive presence						
Triggering Event	1.00	(0.91)	8.72	(8.17)	0.35	(0.30)
Exploration	1.84	(0.98)	13.49	(7.74)	0.56	(0.27)
Integration	1.89	(0.98)	12.69	(7.40)	0.56	(0.27)
Resolution	2.03	(0.86)	12.78	(7.15)	0.56	(0.26)
Other	2.49	(1.26)	12.99	(7.62)	0.57	(0.29)
ICAP modes of cognitive engagement						
Active	1.61	(1.05)	12.94	(7.94)	0.53	(0.30)
Affirmation	2.83	(0.99)	14.97	(6.92)	0.66	(0.26)
Constructive	1.26	(0.89)	10.35	(8.01)	0.45	(0.30)
Interactive	2.32	(0.81)	13.94	(7.15)	0.59	(0.25)

4.2 Experiment 2

In Experiment 2, we addressed RQ2 and looked at the explanatory value of the framework labels in comparison with the dialogue attributes used in Experiment 1.

4.2.1 Predictive performance metrics

Using the same approach as Experiment 1, we trained models using 10-fold cross-validation and compared the Cohen's κ scores for the purpose of model selection (Table 14). Comparing the cross-validation results for Model CoI-2a and Model CoI-2b against Model CoI-1b from Experiment 1 suggested that using the ICAP modes of cognitive engagement as additional features improved the ability of a model to label the CoI phases of cognitive presence (Cohen's $\kappa = 0.460$), but that using the extended set of ICAP modes was no better than using the basic ICAP modes. Thus, for data that has already been labelled with the ICAP modes of cognitive engagement, there would be no benefit in revising those labels to be finer-grained before using an automated system to add labels for the CoI phases of cognitive presence.

Table 14. Cross-validation results for Experiment 2: outcome metrics and the best value for the `mtry` tuning parameter. The best results for each framework are in bold.

Model	Outcome classes	Best `mtry`	Macro F_1	Cohen's κ
CoI phases of cognitive presence				
Model CoI-2a	5 phases, rebalanced with SMOTE	35	**0.553**	**0.460**
Model CoI-2b	5 phases, rebalanced with SMOTE	57	0.551	0.459
ICAP modes of cognitive engagement				
Model ICAP-4	4 modes, rebalanced with SMOTE	35	**0.710**	**0.657**
Model ICAP-6	6 extended modes, rebalanced with SMOTE	57	0.587	0.590

Similarly, the cross-validation results for the model that was trained to label new data with the ICAP modes of cognitive engagement using the CoI phases of cognitive presence as features (Model ICAP-4) showed 'substantial' agreement with the gold-standard labels (Cohen's $\kappa = 0.657$), and an improvement over Model ICAP-1b from Experiment 1. Additionally, Model ICAP-6, which was trained to label new data with the extended set of ICAP modes of cognitive engagement, achieved 'moderate' agreement (Cohen's $\kappa = 0.590$) with the gold-standard labels.

On the basis of the cross-validation results, we used Model CoI-2a and Model ICAP-4 respectively to assign labels to the held-out data from the final offering of the course in order to test the generalisability of the models. The outcome metrics on the test set are shown in Table 15. Labelling the CoI phases of cognitive presence in the test data using Model CoI-2a, we found 'fair' agreement (Cohen's $\kappa = 0.404$) with the gold-standard human coding. The macro-averaged F_1 score was 0.546. These results are higher than the equivalent measures for Model CoI-1b in Experiment 1 (Table 10), indicating that the addition of the ICAP modes of cognitive engagement as model features improved the model. In fact, the F_1 scores for every class of the outcome variable showed an improvement over the earlier results. The only individual measure that decreased was the Precision for the *Triggering Event* class.

Meanwhile, Model ICAP-4 obtained 'substantial' agreement (Cohen's $\kappa = 0.695$) with the gold-standard labels for the

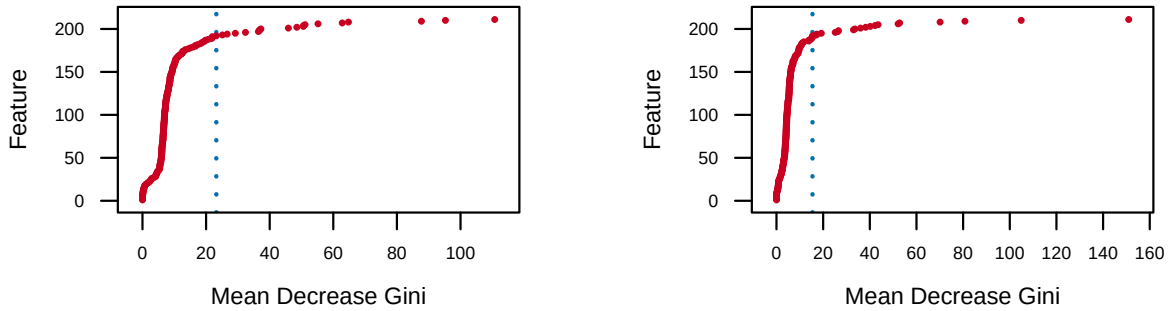
Table 15. Experiment 2: Outcome metrics on the held-out test data.

Outcome variable	Precision	Recall	F_1	Macro F_1	Cohen's κ
CoI phases of cognitive presence, using Model CoI-2a				0.546	0.404
Triggering Event	0.694	0.893	0.781		
Exploration	0.488	0.553	0.519		
Integration	0.561	0.554	0.558		
Resolution	0.429	0.136	0.207		
Other	0.692	0.643	0.667		
ICAP modes of cognitive engagement, using Model ICAP-4				0.765	0.695
Active	0.839	0.553	0.667		
Affirmation	0.615	0.889	0.727		
Constructive	0.838	0.830	0.834		
Interactive	0.767	0.904	0.830		

standard ICAP modes of cognitive engagement, with a macro-averaged F_1 score of 0.765. However, these scores were broadly the same as the results from Experiment 1. Precision for both *Interactive* and *Constructive* modes decreased while their Recall increased. Recall for *Active* mode also decreased but its Precision increased. The small improvements in the individual F_1 scores for *Interactive*, *Constructive*, and *Affirmation* modes were offset by the lower score for *Active* mode to leave the overall macro-averaged F_1 and Cohen's κ scores virtually unchanged from the results in Experiment 1. We conclude that adding the CoI phases of cognitive presence as model features did not improve the ability of the model to label the ICAP modes of cognitive engagement on new data. Nevertheless, it is instructive to observe how the different framework constructs correlated with the outcome classes, and how their discriminatory power compared to those of the dialogue attributes.

4.2.2 Analysis of variable importance

The distribution of the Mean Decrease Gini (MDG) values across the model features are shown in Figure 6. The top 20 features by importance for each framework are listed in Tables 16 and 17. We note that the relationships between the dialogue attributes and the classes of the outcome variable for both models were the same as in Experiment 1, because this is the same data set. Figure 10 in the Appendix shows the distributions for the important features that were not included in the top 20 lists in Experiment 1. Our main objective in this second experiment was to examine how informative the framework constructs were, and to compare their MDG scores with those of the dialogue attributes used in both experiments, in order to answer RQ2.



(a) Model CoI-2a **(b)** Model ICAP-4

Figure 6. Mean Decrease Gini indicating variable importance in the best models for each framework in Experiment 2. In both cases, the vertical dotted line separates the top 20 features.

For both frameworks, many of the same dialogue attributes appeared in the top 20 as in Experiment 1, in approximately the same order. The one-hot attributes generated from the framework constructs were interspersed among them according to their relative importance in the model. Let us now look at each of the frameworks in turn.

The binary attributes generated from all four of the ICAP modes of cognitive engagement appear in the top 20 features for Model CoI-2a, which labels the CoI phases of cognitive presence (Table 16). The discriminative power of the `ICAP-I` feature was second only to the number of words in the message. The mean value for this attribute increased with deeper CoI phases of cognitive presence (Table 18). The `ICAP-F` attribute is third on the list and strongly indicated the *Other* class. `ICAP-A` came in at position 6, with the highest values being seen for messages in the *Triggering Event* and *Other* classes. `ICAP-C` is at position 20 in the table. The values of this attribute for the *Resolution* and *Other* classes were lower than for the other classes, although

Table 16. The 20 most important features in Model CoI-2a, the best CoI cognitive presence model for Experiment 2, ranked from most to least important by Mean Decrease Gini (MDG). The rank position of each feature in Experiment 1 is also shown. Features that appear in the top 20 for the best ICAP model (Table 17) are shown with a coloured background.

Rank	Feature	Description	Previous Rank	MDG
1	`cm.DESWC`	Number of words	1	110.76
2	`ICAP-I`	Labelled as *Interactive*	-	95.25
3	`ICAP-F`	Labelled as *Affirmation*	-	87.71
4	`message.is.first`	First message	2	64.76
5	`message.depth`	Message depth in discussion	5	62.81
6	`ICAP-A`	Labelled as *Active*	-	55.19
7	`cm.LDTTRa`	Lexical diversity, all words	6	51.11
8	`liwc.SemiC`	Number of semi-colons	7	50.66
9	`liwc.posemo`	Number of +ve emotion words	3	50.59
10	`cm.WRDMEAc`	Meaningfulness	4	48.53
11	`message.replies.direct`	Number of direct replies	10	45.89
12	`cm.WRDHYPn`	Hypernyms for nouns	9	37.22
13	`liwc.QMark`	Number of question marks	8	36.85
14	`message.replies.all`	Total number of replies	15	36.76
15	`liwc.discrep`	Number of discrepancy words	12	36.45
16	`liwc.affect`	Number of affective process words	11	32.38
17	`liwc.money`	Number of money words	13	29.20
18	`message.thread.size`	Discussion size	14	26.63
19	`cm.DESWLltd`	SD of word length in letters	17	25.12
20	`ICAP-C`	Labelled as *Constructive*	-	23.21

Table 17. The 20 most important features in Model ICAP-4, the best ICAP model for Experiment 2, ranked from most to least important by Mean Decrease Gini (MDG). The rank position of each feature in Experiment 1 is also shown. Features that appear in the top 20 for the best CoI model (Table 16) are shown with a coloured background.

Rank	Feature	Description	Previous Rank	MDG
1	`message.depth`	Message depth in discussion	1	150.99
2	`CoI-triggering`	Labelled as *Triggering Event*	-	104.89
3	`liwc.assent`	Number of expressions of assent	2	80.78
4	`message.replies.direct`	Number of direct replies	3	70.14
5	`message.replies.all`	Total number of replies	4	52.77
6	`cm.DESWC`	Number of words	5	52.12
7	`cm.LDVOCD`	Lexical diversity, VOCD	6	43.53
8	`cm.LSAGN`	LSA given-new ratio	9	42.24
9	`cm.LDTTRa`	Lexical diversity, all words	7	40.19
10	`CoI-other`	Labelled as *Other*	-	38.23
11	`liwc.posemo`	Number of +ve emotion words	8	36.03
12	`cm.DESPL`	Mean length of paragraphs	10	33.62
13	`cm.DESSC`	Number of sentences	11	33.01
14	`liwc.affect`	Number of affective process words	13	26.58
15	`CoI-exploration`	Labelled as *Exploration*	-	26.29
16	`liwc.QMark`	Number of question marks	12	25.17
17	`cm.LDTTRc`	Lexical diversity, content words	14	19.22
18	`cm.DESWLltd`	SD of word length in letters	19	17.26
19	`liwc.Period`	Number of periods	15	17.13
20	`cm.DESWLsyd`	SD of the mean number of syllables	-	15.43

the standard deviation was relatively high across all classes. In contrast, `ICAP-O`, indicating *Off-task* messages, appeared near the end of the list, at position 201 out of 211. From this we see that the *Interactive* and *Affirmation* labels in particular provide useful information that can help a model to assign the correct CoI phases of cognitive presence to new messages. However, the relationship between the two frameworks is not a simple one and other evidence will also be required.

Table 18. Mean and standard deviations of the one-hot attributes generated from the CoI phases of cognitive presence in Experiment 2, for each of the ICAP modes of cognitive engagement.

	`ICAP-A`		`ICAP-F`		`ICAP-C`		`ICAP-I`		`ICAP-O`	
	mean	(SD)	mean	(SD)	mean	(SD)	mean	(SD)	mean	(SD)
Triggering Event	0.53	(0.50)	0.00	(0.00)	0.47	(0.50)	0.01	(0.08)	0.00	(0.00)
Exploration	0.19	(0.39)	0.03	(0.17)	0.48	(0.50)	0.30	(0.46)	0.00	(0.04)
Integration	0.02	(0.15)	0.01	(0.12)	0.41	(0.49)	0.56	(0.50)	0.00	(0.04)
Resolution	0.01	(0.10)	0.01	(0.10)	0.22	(0.42)	0.76	(0.43)	0.00	(0.00)
Other	0.41	(0.49)	0.34	(0.47)	0.14	(0.35)	0.04	(0.20)	0.06	(0.25)

In Model ICAP-4, three of the five binary attributes derived from the CoI phases of cognitive presence were among the top 20 features (Table 17). `CoI-triggering` came in at position 2, with the highest mean values being seen for the *Active* mode, followed by *Constructive* (Table 19). At position 10, we found `CoI-other`, strongly indicative of *Affirmation* messages. `CoI-exploration`, at position 15, had a smoother distribution and was most likely to indicate a *Constructive* message. The attributes indicating *Integration* and *Resolution* did not appear in the top 20 list: `CoI-integration` was number 29, while `CoI-resolution` was number 196 out of 211. Sixteen of the remaining top 20 attributes also appeared in the top 20 list in Experiment 1. The standard deviation of the mean number of syllables (`cm.DESWLsyd`) was a new addition at position 20; the highest values were likely to indicate *Constructive* mode.

Table 19. Mean and standard deviations of the one-hot attributes generated from the ICAP modes of cognitive engagement in Experiment 2, for each of the CoI phases of cognitive presence.

	`CoI-triggering`		`CoI-exploration`		`CoI-integration`		`CoI-resolution`		`CoI-other`	
	mean	(SD)	mean	(SD)	mean	(SD)	mean	(SD)	mean	(SD)
Active	0.45	(0.50)	0.35	(0.48)	0.03	(0.18)	0.00	(0.05)	0.16	(0.37)
Affirmation	0.00	(0.00)	0.27	(0.45)	0.09	(0.29)	0.01	(0.12)	0.63	(0.49)
Constructive	0.20	(0.40)	0.45	(0.50)	0.29	(0.45)	0.03	(0.18)	0.03	(0.16)
Interactive	0.00	(0.06)	0.36	(0.48)	0.49	(0.50)	0.14	(0.35)	0.01	(0.10)

We have already seen that adding model features derived from the CoI phases of cognitive presence did not improve the overall predictive power of the model. The MDG values allowed us to see how much information each of these features provided independently. It seems that the *Other* label could be used to identify *Affirmation* messages somewhat reliably, while a *Triggering Event* message is highly unlikely to be labelled as *Interactive* or *Affirmation*.

5. Discussion

In this section, we start by looking at the findings that relate directly to each of our research questions, then move on to more general findings. Next, we consider the limitations of the present study and how the results compare with prior work. We conclude by outlining implications for research and practice.

5.1 Automatic labelling using dialogue attributes works better for ICAP than for CoI

RQ1 asked about the relationship between the dialogue attributes and the framework labels. We addressed the question by training random forest models for each framework using the dialogue attributes as features. There was a notable disparity in the predictive power of the models we trained for the two frameworks. The best model for labelling the CoI phases of cognitive presence (Model CoI-2a) achieved only Cohen's $\kappa = 0.404$ on the held-out test data from the final course session, whereas the human annotators who created the gold standard reported Cohen's $\kappa = 0.974$, indicating that there is substantial room for improvement. We particularly noted the low F_1 score for the *Resolution* phase, for which there were only 85 training examples. In previous work, the *Resolution* phase has sometimes been combined with the *Integration* phase to create a 'higher-order thinking' category containing more examples (McKlin, 2004; Schrire, 2006).

In contrast, the models that were trained to label the ICAP modes of cognitive engagement achieved similar results in both experiments (Cohen's $\kappa = 0.695$ in Experiment 2). This result is better than the reported inter-annotator agreement (Cohen's κ

= 0.623). However, our experiments used only 4 ICAP modes, rather than the 6 labels of the extended ICAP taxonomy that were used to assess human agreement. Although there were only 66 examples for the *Affirmation* mode in the training data, its F_1 score was only a little lower than for the other modes.

One possible explanation for the difference in performance between the two frameworks could be that there are fewer distinct classes in the outcome variable for ICAP, making it easier for a model to choose the correct one. However, when we trained a model to predict all 6 modes in the extended ICAP taxonomy (Model ICAP-6), the results from cross-validation (Cohen's $\kappa = 0.590$) were still better than any we saw for the 5 CoI phases of cognitive presence. We therefore propose that a better explanation is that the ICAP modes are more closely related to the linguistic attributes of the messages than are the CoI phases of cognitive presence. This seems reasonable, since the definition of the ICAP framework (Chi & Wylie, 2014) emphasises its focus on overt, observable, behaviours as proxies for the knowledge change processes that constitute learning. For example, the definitions for both *Interactive* and *Constructive Reasoning* modes look for "*explanation or reasoning about the current topic*" – behaviours which correspond directly to dialogue-level attributes. The label definition for the *Affirmation* mode is also stated in terms of dialogue-level features, with the number of expressions of assent being especially predictive.

In this study, the only context that was provided to the models related to the structure of the discussion and the position of a message within that structure. Additional information relating to the content of previous messages, such as textual similarity measures (Kovanović et al., 2016; Atapattu et al., 2019), could thus be expected to improve the models' ability to distinguish between labels. For example, some students were observed to respond to clarification questions by repeating the question in full before giving a brief answer. Relying on basic metrics, such as sentence length, would treat such a message the same as another where a longer answer is given without repeating the question. However, as the discussion grows, there could be technical limitations on the amount of dialogue history that can reasonably be processed and a recency threshold might need to be introduced. Future work might also consider whether the amount of context that is useful differs between the frameworks.

5.2 There is an asymmetric relationship between the framework labels

RQ2 asked about the explanatory value of the labels from one framework when modelling the other. We hypothesised that the two frameworks would generally provide complementary views on the learning experience, rather than being closely aligned (Section 2.4), with higher-level constructs more likely to be correlated. In Experiment 2 we examined the relationship between them directly by using each of them in turn as input to a model trained to label the constructs from the other. We found that the information provided by the ICAP modes of cognitive engagement led to a small improvement in the outcome metrics for Model CoI-2a that was trained to label the CoI phases of cognitive presence. We saw increases in the F_1 score for each of the individual phases, with macro-averaged F_1 scores improving from 0.515 to 0.546, and Cohen's κ increasing from 0.358 to 0.404 on the held-out test data (Tables 10 and 15). Nevertheless, none of the ICAP modes of cognitive engagement was a direct analogue of any of the CoI phases of cognitive presence. The attributes indicating the *Interactive* and *Affirmation* modes ranked highest as predictive features, while the *Active* mode was more explanatory than the *Constructive* mode. We suggest that the first message effect (Section 5.3) is also relevant here. The closest relationship was between the *Affirmation* mode and the *Other* phase; while messages labelled as *Interactive* were distributed across the *Exploration*, *Integration*, and *Resolution* phases in increasing proportions (Table 18).

There was no similar model improvement in the reverse direction. The Cohen's κ and macro-averaged F_1 scores for Model ICAP-4 were virtually unchanged by the addition of features based on the CoI phases of cognitive presence. Messages that were identified as *Triggering Event*s were split between the *Constructive* and *Active* modes, while *Exploration* messages were relatively evenly spread across all four ICAP modes (Table 19).

Building on the evidence in Section 5.1 of the correspondence between dialogue-level attributes and the ICAP modes of cognitive engagement, we can add here that those attributes might be reasonable proxies for the quality measure defined by the ICAP framework, since features derived from the CoI phases of cognitive presence do not improve the predictive power of such a model. In contrast, the CoI phases of cognitive presence are not well predicted using dialogue attributes alone. We conclude that the frameworks measure different aspects of the quality of student participation.

5.3 Messages that are nested deeper in threads tend to be higher quality

We noted in Section 2.3 that online discussions are expected to progress through each of the CoI phases of cognitive presence in order (although not all will do so), while there was no similar expectation of orderly progression through the ICAP modes of cognitive engagement. This contrast in the framework definitions was partially supported by our experimental results. We saw in Experiment 1 that greater message depth, in terms of nesting within a discussion thread, was generally correlated with indications of higher quality contributions in both frameworks, while strictly chronological message ordering did not help the models to distinguish between labels.

For the CoI phases of cognitive presence, messages in each phase from *Triggering Event* to *Resolution* had an increasing average message depth (Table 13). This observation provides support for the temporal progression expected for the CoI phases of cognitive presence – although within a message sub-thread, rather than in the larger discussion. For the ICAP models,

message depth was the single most informative attribute (Table 12). Messages labelled as *Affirmation* had the greatest average depth. Surprisingly, *Active* messages were found at a greater mean depth than *Constructive* messages (Table 13). One reason for this is that the first message in each discussion thread typically introduces a new problem or topic for discussion (*Constructive* mode), skewing the averages. Further work would be needed to separate out the effect of first message label bias.

The distinction noted here between thread depth and chronological ordering corresponds to the difference between adding another message onto an established thread, and adding a new message at a higher level in the discussion – such as a new response to the opening message. We speculate that a message that is posted later in time at high level may well ignore the content of earlier sub-threads, but a message that extends an existing thread is likely to build on what has gone before within that thread.

5.4 *Affirmation* mode messages tend to display a lack of cognitive presence

We observed many similarities between the predictors for the two frameworks. Some are unsurprising: longer messages were correlated with higher cognitive quality in both frameworks, as were greater lexical diversity and fewer question marks. Other relationships are more complex. Messages displaying higher than average numbers of affective process words and expressions of positive emotion tended to cluster in a single class of the outcome variable (*Other* and *Affirmation*, respectively). Messages in those classes also demonstrated low scores for lexical diversity and 'meaningfulness', compared with other messages, and tended to be the shortest messages. The presence of the *Other* label was a strong indicator of the *Affirmation* label, and vice versa. When the *Affirmation* label was used as a predictive feature to label the CoI phases of cognitive presence in Experiment 2, it played the same role for identifying *Other* messages as the first-message flag did for *Triggering Event* messages.

However, there are important differences in the interpretation of these classes. Whereas the *Other* label indicates that no signs of cognitive presence were evident in a given message, messages with the *Affirmation* label may be relabelled later, based on the label of the message to which they were responding (Yogev et al., 2018). By affirming what was said in an earlier message, the student is thus credited with demonstrating some cognitive engagement, albeit not to the same extent as the original contributor (see the description of the relabelling process in Section 3.2 for details). Since interaction with other students is associated with the greatest learning gains, this variant of the ICAP framework rewards conversational moves that foster interactivity by continuing the conversation and opening the way for further elaboration. In contrast, the CoI framework treats messages of affirmation solely as indicators of social presence. However, recent work (Hu et al., 2020) departed from the original CoI framework definitions and instead treated a simple message of agreement (or disagreement) as a *Triggering Event*, and one that gave reasons for agreement as belonging to the *Integration* phase. Our work here demonstrates that it is important not to neglect the social dimension when evaluating the worth of a discussion forum contribution, and we welcome recent work on the automatic detection of social presence (M. Ferreira et al., 2020). Another factor that may be relevant to social presence and could be considered in future work is the number of unique participants in each thread.

5.5 Limitations

Only a single data set was used for this study. Because of the particular discussion task that was set in that course, the first message of every thread followed a similar format and was typically labelled in the same way. There is no reason to suppose that messages from another course would share this property, so caution is needed in interpreting results relating to features derived from message position. Additionally, the *Passive* mode of the ICAP framework was not used at all, because the data set did not include a record of when students read the messages posted by others, and *Off-task* messages were too infrequent to be used in this study.

The random forest approach allowed us to identify dialogue attributes that are effective in discriminating between the classes of the outcome variable. It provides a value for each one independently. However, it may well be the case that several attributes are related in such a way that their values are correlated with one another. We saw this in Table 12, where a closer look at the attributes at positions 10 and 11 revealed that they were in fact measuring exactly the same thing. Using either one of those attributes would be valuable, but there is no additional benefit in using both together. Different analytical methods are needed to measure the marginal value of each attribute. This should be tackled as a priority in future work, since it could provide some much-needed nuance and move us beyond the simple view that "longer is better" when it comes to forum messages.

Our study design did not allow us to compare the overall predictive power of the lexical features compared to the structural features. It could be the case that certain aspects of the discussion activity are more closely related to specific dialogue moves and could thus be identified by lexical features. However, we note that both frameworks aim to capture aspects of learning that are not specific to the learning activities: "*instructional tasks are orthogonal to engagement mode*" (Chi & Wylie, 2014, p. 221). Future work could usefully consider how lexical features that are task-specific might interact with those identified in this work as being predictive of framework labels.

5.6 Comparison with previous studies

Several previous studies have built classifiers to automate the labelling of the CoI phases of cognitive presence (Section 2.1).

The current work used the same data and the same set of 106 linguistic features from Coh-Metrix and 91 from LIWC that were used in both Kovanović et al. (2016) and the later replication study (Farrow et al., 2019), along with some of the same structural features: message depth within a thread, first and last message indicators, and the number of replies a message received[2]. The Kovanović et al. study reported Cohen's $\kappa = 0.63$, but the replication study demonstrated problems with the way the training and testing data had been prepared and showed that a more realistic result was Cohen's $\kappa = 0.38$, with macro-averaged $F_1 = 0.54$.

The small reduction in the outcome metrics between the replication study and Experiment 1 in the current work (Cohen's $\kappa = 0.358$, macro-averaged $F_1 = 0.515$ (Table 10)) reflects the additional predictive value of the structural features from the two earlier studies that were not used in Experiment 1. Those features were cosine similarity to the previous and next message; internal coherence across the sentences within a message; and the count of relevant named entities in the message. Similarity measures were not used in the current work primarily because the definitions of *next* and *previous* message were deemed to be unclear in the context of a threaded forum where new messages could be added at any level. The other two measures were excluded from the current study because they relied on external resources which would not be portable across domains.

Two studies looking at discussion forum messages written in Portuguese (Neto et al., 2018; Barbosa et al., 2020) also used a collection of linguistic and structural features similar to that in the current work. Fewer linguistic features were available for Portuguese than for English in both Coh-Metrix and LIWC. The Portuguese-only study (Neto et al., 2018) used 48 features from Coh-Metrix and 24 adapted from LIWC, while the cross-language English-Portuguese study (Barbosa et al., 2020) included 38 from Coh-Metrix and 64 from LIWC. Both studies additionally used message depth, number of replies, cosine similarity to the previous and next message, and the count of named entities[3]. The features used in the current work and related previous studies are summarised in Table 21 in the Appendix. In the Portuguese-only study, the data was split into training and testing sets using stratified random sampling. The classifier scored Cohen's $\kappa = 0.72$ on the test data (macro-averaged $F_1 = 0.63$), higher on both measures than the comparable results in the current work. This type of random sampling works well when all the data points are independent; but where messages from the same thread are assigned to different partitions, the results can be misleading; such classifiers may not perform well on new data (Farrow et al., 2019). The data in the cross-language study did not need to be split: the classifier was trained on the same English-language data as the current study and evaluated on Portuguese-language data. The reported result was Cohen's $\kappa = 0.53$. However, none of the *Integration* and *Resolution* messages in the test data were classified correctly, and the confusion matrix presented in that work (Barbosa et al., 2020, Table 4c) corresponds to a result of Cohen's $\kappa = 0.44$ with macro-averaged $F_1 = 0.38$. The current study thus scored lower on Cohen's κ but higher on the macro-averaged F_1 measure which gives equal weight to every class of the outcome variable.

Comparing the top 20 most predictive features from Experiment 1 (Table 11) with those in the top 20 reported by Kovanović et al., 8 of a possible 18 common features appeared in both lists. The replication study (Farrow et al., 2019) did not report which features were most predictive in the revised model. In fact, only 9 of the top 20 features were shared between the original study and the replication, whereas 15 of 18 common features from Experiment 1 featured in the top 20 of the replication study. It is less straightforward to compare the top features with those from the papers using Portuguese data (Neto et al., 2018; Barbosa et al., 2020) since the reduced feature sets available in the Portuguese versions of Coh-Metrix and LIWC are not listed in full. Features that scored highly in the current study may not have been available to those classifiers (for example, `cm.WRDMEAc`, defined as 'meaningfulness'). Our best estimate is that 6 from a possible 12 common features from Experiment 1 are also in the top 20 for the Portuguese-only study (Neto et al., 2018), while 3 from 11 common features appear in the top 20 for the cross-language study (Barbosa et al., 2020). The comparative rankings of features are summarised in Table 22 in the Appendix. Three features appeared in the top 20 across all five studies: the number of words in the message, the message depth in the discussion, and the number of question marks.

Previous studies where the ICAP modes of cognitive engagement were labelled automatically by a classifier (Section 2.2) have most commonly used a bag-of-words approach (Wang et al., 2015; Atapattu et al., 2019). The study of comments made in an electronic textbook (Yogev et al., 2018) additionally used several context features, including cosine similarity between the student comment and the highlighted text. None of these studies used linguistic features from Coh-Metrix or LIWC, and the data sets themselves were very different, so a direct comparison of the results with the current study is not feasible in the way we did for the CoI phases of cognitive presence.

5.7 Implications for research and practice

Conceptual frameworks like CoI and ICAP have generally been used for post-hoc analysis of discussion forum data, as in the present study (Schrire, 2006; Garrison, 2016; Kovanović et al., 2016; Neto et al., 2018; Farrow et al., 2019). However, as automated classifiers improve, it becomes more feasible to deploy them inside a learning analytics system while a course is in progress. Automatically generating framework labels in real-time or with a short delay can allow instructors to get a high-level

[2] In the current work, both direct and indirect replies were counted, and additional structural features were added (Table 4).

[3] All named entities were included in the count, not only those relevant to the discussion domain.

overview of the discussion quality while there is time to intervene (Yogev et al., 2018).

One outcome of this study is a clearer picture of the explanatory power of dialogue attributes and the extent to which they could be used as a low-cost proxy for the framework labels. We saw, for example, that message depth in the reply-based structure of message threads was a better predictor of cognitive quality than chronological order across both frameworks. This was particularly the case for ICAP, where message depth was the most explanatory feature in the model (Table 12). We therefore recommend that users and providers of discussion boards should ensure that information about the threaded structure of the discussion is always preserved and made available for analysis.

We also saw that messages containing more question marks tended to be correlated with lower cognitive quality in both frameworks, and that this feature was ranked highly in terms of explanatory power in the predictive models. We note that it is often the case that participation instructions for graded discussion forums specify expected engagement in terms of asking a certain number of questions (Gilbert, 2002; Gašević et al., 2015), and caution that this may encourage the production of many shallow sub-threads, since it is common for students to do the minimum that is required (Vellukunnel et al., 2017). To overcome this problem, we suggest that educators should explicitly reward students for generating deeper discussion threads and building on what has been said by others, rather than simply asking questions. One simple approach would be to limit the number of replies to the original message that qualify for participation credit. After the first two (for example) top-level responses, students seeking credit would need to expand one of those threads further. Even when one of the early responses lacked substance, a follow-up message could reference it while steering the discussion in a more desirable direction, generating higher quality discussion overall. Alternative approaches based on some kind of group-wide incentive are probably too complex for most discussion-based learning activities.

We investigated the relationship between the frameworks directly in order to answer RQ2 by considering a data set that already had labels corresponding to the constructs from one framework and looking at how similar labels from the other could be added automatically (Experiment 2). We found that using the ICAP modes of cognitive engagement produced a small improvement in the predictive performance metrics for the model that was trained to label the CoI phases of cognitive presence, but there was no improvement at all in the other direction. This indicated that the frameworks measure different aspects of the quality of student participation. We conclude that researchers who use both frameworks together would therefore gain richer insights. For example, an intervention study could use both measures, manually or automatically labelled, to assess the impact of the treatment – perhaps capturing changes in contributions from *Active* to *Constructive* mode even if they did not progress from *Exploration* to *Integration*. If a single quality measure was needed, some mechanism would need to be devised to combine them, and that might differ depending on the goal of the learning activity. If the focus was on constructive knowledge building, then the chosen quality measure might prioritise the CoI phases of cognitive presence; but if it was defined more broadly, perhaps the ICAP modes of cognitive engagement would be a better choice as the primary measure. Our results in this study did not provide any examples of one framework-based quality measure increasing while the other decreased.

We paid particular attention to the treatment of messages of affirmation, such as thanks and agreement. In previous work these messages were often treated as being of low worth, but recent work has started to reconsider the role they play in encouraging interaction (Yogev et al., 2018; Hu et al., 2020). Many social media platforms now provide non-textual ways of indicating agreement and affirmation, such as a 'thumbs up' or 'heart' symbol. If the platforms that host course discussion boards adopt this approach, then it will be important to preserve such indicators for analysis. We note that composing a written message, however brief, requires greater effort compared to simply clicking a symbol; it also provides scope for personalisation and elaboration. Future research should therefore compare the two types of affirmation, text and symbols, to discover what effect the modality has on how the affirmation is perceived and how the discussion develops.

6. Conclusion and Future Work

Our aim was to identify dialogue attributes that could be used to discriminate between discussion contributions of varying quality. Our expectation was that the CoI and ICAP frameworks would provide complementary perspectives on how we might begin to quantify students' cognitive engagement with the intellectual content of a course through discussion forum messages. We also expected to identify specific dialogue attributes that could be used directly to guide both instructors and students to improve the quality of online discussions.

We trained several random forest models to label the constructs from each framework in new data using as inputs linguistically-motivated dialogue attributes as well as structural features of the discussion. We found that several simple measures of contribution size, such as the number of words in a message, were correlated with greater quality in both frameworks; while other correlations were framework-specific, such as the higher numbers of personal pronouns found in the *Active* mode in ICAP. We hypothesise that the dialogue attributes that were identified by both frameworks will be generally useful as proxies for the quality of student discussion contributions across a broad range of learning situations.

Comparing the two frameworks directly on the same data set, we found that using the constructs from one framework as additional input while labelling the constructs from the other framework produced little to no improvement in the predictive

performance metrics. These results indicated that the frameworks are not closely aligned and suggests that instead, they measure different aspects of quality. The attributes that were correlated with quality measures in only one framework might therefore be relevant in a more limited set of learning contexts. Researchers could gain richer insights by using both frameworks together in future to assess message quality – an approach that is likely to become increasingly feasible, thanks to the development of automated classifiers for both CoI (Kovanović et al., 2016; Neto et al., 2018; Farrow et al., 2019; M. Ferreira et al., 2020; Hu et al., 2020) and ICAP (Yogev et al., 2018; Atapattu et al., 2019).

We also considered the different treatment of affirmations in the two frameworks and in prior work. In CoI, they are generally considered solely as indicators of social presence, with no value in terms of cognitive presence. In contrast, in studies using ICAP, their value can depend on the content of the earlier message they are affirming, due to the greater value placed on interaction as a deeper mode of cognitive engagement. Future work incorporating the automatic detection of social presence alongside cognitive presence would allow researchers and instructors to take this important aspect of learning into account more easily when using the CoI framework.

We successfully identified a small set of dialogue attributes that were highly predictive of quality according to both frameworks. Based on these, we proposed a modification to common participation requirements to encourage students to generate deeper threads rather than more top-level questions, since the latter typically demonstrate little connection with other parts of the discussion. While contribution quantity is also highly correlated with measures of participation, simply setting a minimum threshold on message length is unlikely to improve learning and would certainly harm important social exchanges such as affirmations. Future research should look beyond contribution quantity to consider other dialogue attributes that indicate the quality of participation.

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

Atapattu, T., Thilakaratne, M., Vivian, R., & Falkner, K. (2019). Detecting cognitive engagement using word embeddings within an online teacher professional development community. *Computers & Education*. doi: 10.1016/j.compedu.2019.05.020

Barbosa, G., Camelo, R., Cavalcanti, A. P., Miranda, P., Mello, R. F., Kovanović, V., & Gašević, D. (2020). Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 605–614). New York, NY, USA: ACM. doi: 10.1145/3375462.3375496

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: the SOLO taxonomy (structure of the observed learning outcome)*. New York: Academic Press.

Bloom, B. S., Krathwohl, D. R., & B., M. B. (1956). *Taxonomy of educational objectives : the classification of educational goals. Handbook 1, Cognitive domain*. London: Longman Group Ltd.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi: 10.1613/jair.953

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. doi: 10.1207/s15516709cog1302_1

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

Corich, S., Kinshuk, & Hunt, L. M. (2006). Measuring critical thinking within discussion forums using a computerised content analysis tool. In *Proceedings of the 5th International Conference on Networked Learning* (pp. 1–8). Retrieved from `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.8598&rep=rep1&type=pdf`

Corich, S., Kinshuk, & Jeffrey, L. M. (2007). Changing focus from group to individual: Using an automated tool to measure evidence of critical thinking in discussion forums. In *IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2007)* (pp. 163–172).

Cross, S., Waters, Z., Kitto, K., & Zuccon, G. (2017). Classifying help seeking behaviour in online communities. In *Proceedings of the Seventh International Learning Analytics & Knowledge conference* (pp. 419–423). New York, NY, USA: ACM. doi: 10.1145/3027385.3027442

Cui, Y., & Wise, A. F. (2015). Identifying content-related threads in MOOC discussion forums. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (pp. 299–303). New York, NY, USA: ACM. doi: 10.1145/2724660.2728679

Farrow, E., Moore, J., & Gašević, D. (2020). Dialogue attributes that inform depth and quality of participation in course discussion forums. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 129–134). New York, NY, USA: ACM. doi: 10.1145/3375462.3375481

Farrow, E., Moore, J., & Gašević, D. (2019). Analysing discussion forum data: A replication study avoiding data contamination. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 170–179). New York, NY, USA: ACM. doi: 10.1145/3303772.3303779

Farrow, E., Moore, J., & Gašević, D. (2021). A network analytic approach to integrating multiple quality measures for asynchronous online discussions. In *Proceedings of the 11th International Conference on Learning Analytics & Knowledge.* New York, NY, USA: ACM. doi: 10.1145/3448139.3448163

Ferguson, R., Wei, Z., He, Y., & Buckingham Shum, S. (2013). An evaluation of learning analytics to identify exploratory dialogue in online discussions. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 85–93). New York, New York, USA: ACM Press. doi: 10.1145/2460296.2460313

Ferreira, M., Rolim, V., Mello, R. F., Lins, R. D., Chen, G., & Gašević, D. (2020). Towards automatic content analysis of social presence in transcripts of online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 141–150). New York, NY, USA: ACM. doi: 10.1145/3375462.3375495

Ferreira, R., Kovanović, V., Gašević, D., & Rolim, V. (2018). Towards combined network and text analytics of student discourse in online discussions. In *Proceedings of the International Conference on Artificial Intelligence in Education* (Vol. 10331, pp. 111–126). Springer International Publishing. doi: 10.1007/978-3-319-93843-1_9

Gardner, J., Brooks, C., Andres, J. M., & Baker, R. (2018). Replicating MOOC predictive models at scale. In *Proceedings of the 5th Annual ACM Conference on Learning at Scale* (pp. 1–10). doi: 10.1145/3231644.3231656

Garrison, D. R. (2011). *E-learning in the 21st century: A framework for research and practice* (2nd ed.). New York and London: Routledge. doi: 10.4324/9780203838761

Garrison, D. R. (2016). *Thinking collaboratively: Learning in a community of inquiry*. New York and London: Routledge.

Garrison, D. R., Anderson, T., & Archer, W. (2000). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education*, *2*(2), 87–105. doi: 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. doi: 10.1080/08923640109527071

Gašević, D., Adesope, O., Joksimović, S., & Kovanović, V. (2015). Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The Internet and Higher Education*, *24*, 53–65. doi: 10.1016/j.iheduc.2014.09.006

Gilbert, P. K. (2002). *Assessing the impact of the structuredness of online discussion protocols on meaningful discourse* (Unpublished doctoral dissertation). George Mason University.

Harrak, F., Bouchet, F., Luengo, V., & Gillois, P. (2018). Profiling students from their questions in a blended learning environment. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 102–110). New York, NY, USA: ACM. doi: 10.1145/3170358.3170389

Hu, Y., Donald, C., Giacaman, N., & Zhu, Z. (2020). Towards automated analysis of cognitive presence in MOOC discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 135–140). New York, NY, USA: ACM. doi: 10.1145/3375462.3375473

Kovanović, V., Joksimović, S., Gašević, D., & Hatala, M. (2014). Automated cognitive presence detection in online discussion transcripts. *CEUR Workshop Proceedings*, *1137*. Retrieved from http://ceur-ws.org/Vol-1137/LA_machinelearning_submission_1.pdf

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 15–24). New York, NY, USA: ACM. doi: 10.1145/2883851.2883950

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi: 10.2307/2529310

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, p. 9). JMLR.org. Retrieved from http://proceedings.mlr.press/v32/le14.pdf

McKlin, T. (2004). *Analyzing cognitive presence in online courses using an artificial neural network* (Doctoral dissertation, Georgia State University). Retrieved from http://scholarworks.gsu.edu/msit_diss/1

McKlin, T., Harmon, S. W., Evans, W., & Jones, M. G. (2001). Cognitive presence in web-based learning: A content analysis of students' online discussions. In *Annual proceedings of selected research and development and practice papers presented at the National Convention of the Association for Educational Communications and Technology* (Vol. 1-2, pp. 272–277). Retrieved from https://files.eric.ed.gov/fulltext/ED470101.pdf

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

Meyer, K. A. (2004). Evaluating online discussions: four different frames of analysis. *Journal of Asynchronous Learning Networks*, *8*(2), 101–114.

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 *Proceedings of the 13th European Conference on Technology Enhanced Learning* (pp. 245–261). Cham: Springer International Publishing. doi: 10.1007/978-3-319-98572-5_19

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

Taskin, Y., Hecking, T., Hoppe, H. U., Dimitrova, V., & Mitrovic, A. (2019). Characterizing comment types and levels of engagement in video-based learning as a basis for adaptive nudging. In *Proceedings of the 14th European Conference on Technology Enhanced Learning* (pp. 362–376). Cham: Springer International Publishing. doi: 10.1007/978-3-030-29736-7_27

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. doi: 10.1177/0261927X09351676

Vellukunnel, M., Buffum, P., Boyer, K. E., Forbes, J., Heckman, S., & Mayer-Patel, K. (2017). Deconstructing the discussion forum: Student questions and computer science learning. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education - SIGCSE '17* (pp. 603–608). New York, NY, USA: ACM. doi: 10.1145/3017680.3017745

Wang, X., Wen, M., & Rosé, C. P. (2016). Towards triggering higher-order thinking behaviors in MOOCs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 398–407). New York, NY, USA: ACM. doi: 10.1145/2883851.2883964

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains. In *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 226–233). Massachusetts, USA: International Educational Data Mining Society. Retrieved from http://www.educationaldatamining.org/EDM2015/proceedings/full226-233.pdf

Waters, Z., Kovanović, V., Kitto, K., & Gašević, D. (2015). Structure matters: Adoption of structured classification approach in the context of cognitive presence classification. In *Information Retrieval Technology* (Vol. 9460, pp. 227–238). Cham: Springer International Publishing. doi: 10.1007/978-3-319-28940-3_18

Wise, A. F., & Cui, Y. (2018). Unpacking the relationship between discussion forum participation and learning in MOOCs. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 330–339). New York, New York, USA: ACM Press. doi: 10.1145/3170358.3170403

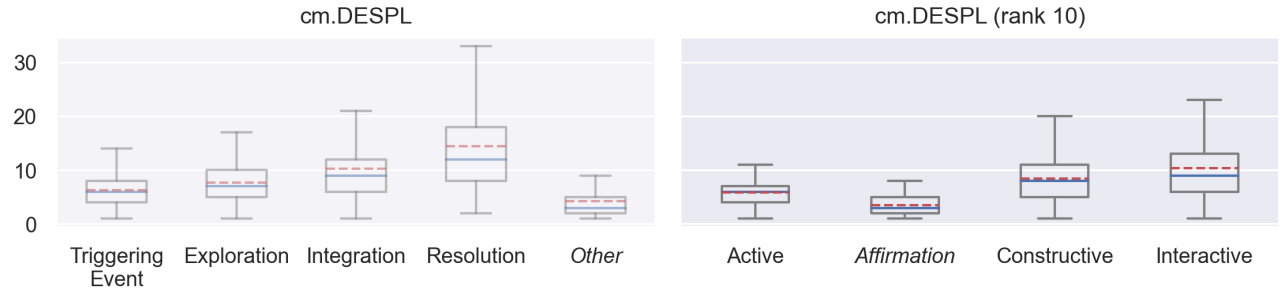
Wise, A. F., Cui, Y., & Vytasek, J. (2016). Bringing order to chaos in MOOC discussion forums with content-related thread identification. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 188–197). New York, New York, USA: ACM Press. doi: 10.1145/2883851.2883916

Yogev, E., Gal, K., Karger, D., Facciotti, M. T., & Igo, M. (2018). Classifying and visualizing students' cognitive engagement in course readings. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (pp. 1–10). New York, NY, USA: ACM. doi: 10.1145/3231644.3231648

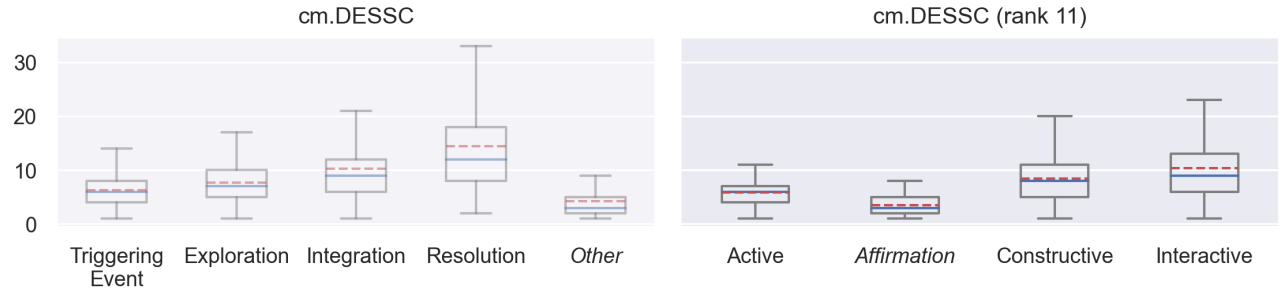
Appendix

Table 20. Redacted sample messages from the data set with each framework label.

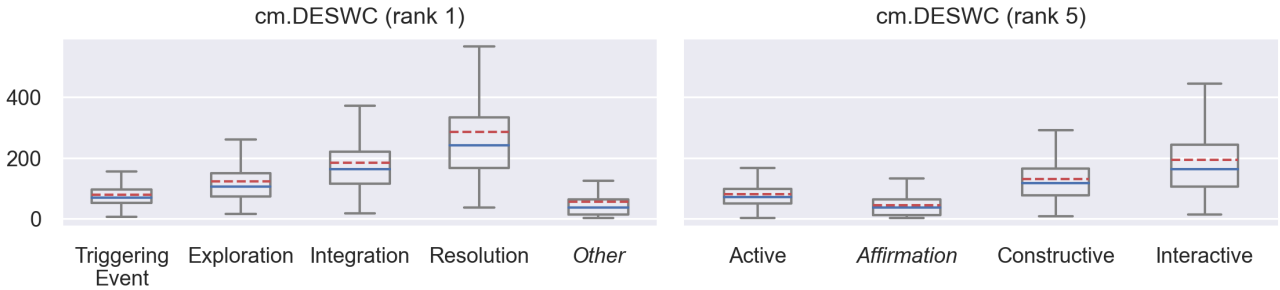
Label	Sample message (redacted)
CoI phases of cognitive presence	
Triggering Event	Hi [NAME] Thank very much for your presentation. It was very informative and interesting! Just a question: Can the concept of [X] and the [X] software that you demonstrated in your presentation also be used for [Y] and [Z]? Or does this originate from [Z]/ [Y]? Please let me know. Thanks in advance [NAME]
Exploration	Hi The paper states that they do not believe there is much difference between an [X] and a [Y] and that actually in the last few years especially in the corporate world there is normally not a title of '[Y]' rather most folks in the field are 'bundled' in the [X] title. None-the-less it is certainly possible given the small population size that the folks who participated in the surveys did not include a good enough 'cut' of the industry roles which would serve to skew results. Sounds like you and I are very skeptical of the results.. cannot blame you. Regards [NAME]
Integration	Hello [NAME] Thanks very much for your feedback. I agree with you that [X] would provide better [A] and [B] but [X] requires more resources [Y] provides a quick and easy solution. My idea of implementing a [Y] 4 between the [P] and [Q] mainly to protect important data. There may be some data on the database that you only want to be accessed from certain locations or computers. Best regards [NAME]
Resolution	Hi [NAME] Thanks for the comments. The concept of [X] is to model complexity in a different way. As you have mentioned modeling the viewpoints of [M] who each see from a different perspective is both exciting and daunting. [Y] is an iffy process at best. This is because one persons viewpoint is different than another persons. The role of a [P] is to gather these perspectives then merge them (somehow) into a cohesive model. This is where [X] tries to step in. Eventually when modeling meets code generation fully we will then have automated processes that will generate code for every change whether that is a user role or another related mechanism. At this point [X] is still a work in progress; a research level tool that requires maturation. I do see a future for it although we may not recognize it as such when it arrives. Thanks
Other	Hi [NAME] Thanks for watching. There is so much material to cover in this course that it is very difficult. But it is all very interesting. Looking forward to your question. Cheers [NAME]
ICAP modes of cognitive engagement	
Active General	Hi [NAME]] Like others I had no problem understanding you. Interesting topic. Do you know of any other system which tags video with which to compare [X] to? Cheers [NAME]
Active Targeted	Good presentation although a bit long. In the presentation mention [X]s as a mechanism on one slide and yet mention that [Y] [X]s are akin to stakeholders. Could you explain this a bit more?
Affirmation	Hi [NAME] Thanks for answering my question. The difference is much clearer! [NAME]
Constructive Extending	Hi [NAME] Good job on the presentation. Here are my comments/questions: 1. Do the percentages in the charts differ for non [X] software? If so why? 2. Traditionally requirements have been difficult to nail down upfront. Why would obtaining usability requirements have more success than obtaining business requirements? Thanks [NAME]
Constructive Reasoning	[NAME] Great presentation clear voice. I love your accent. My question – did the authors talk about requirements engineering in the context of various [X] methodologies? Because each [X] treats the requirements engineering process very differently ([Y] versus [Z]) I'm wondering how the topics they covered would change in these contexts. Or am I off track and it would not matter?
Interactive	Hi [NAME] Interesting. You say that since [X] is 40 years old there should be some laws now. But I see it this way; since [X] is 40 years old and there are no laws yet that must mean that they are really hard to define. Which way do you see it? Cheers [NAME]
Off-task	I remember - it just sounds too much with [X] and then its all I can think of is that song [NAME] PS thanks for the post [NAME]



(a) Mean length of paragraphs: Coh-Metrix label `DESPL`

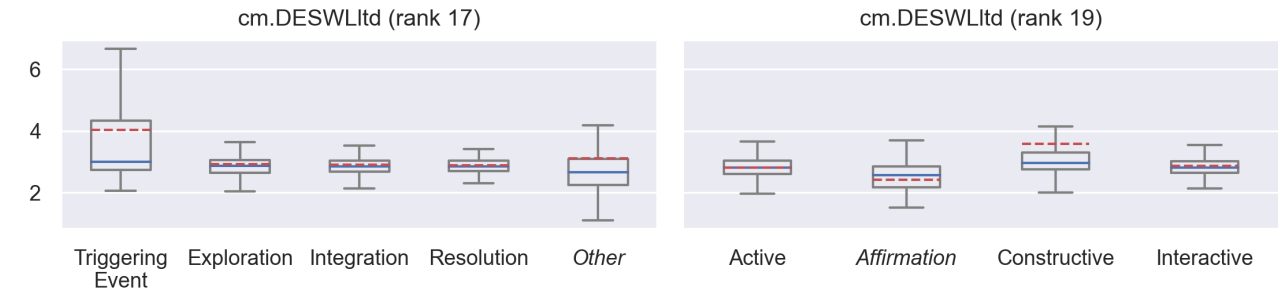


(b) Number of sentences: Coh-Metrix label `DESSC`

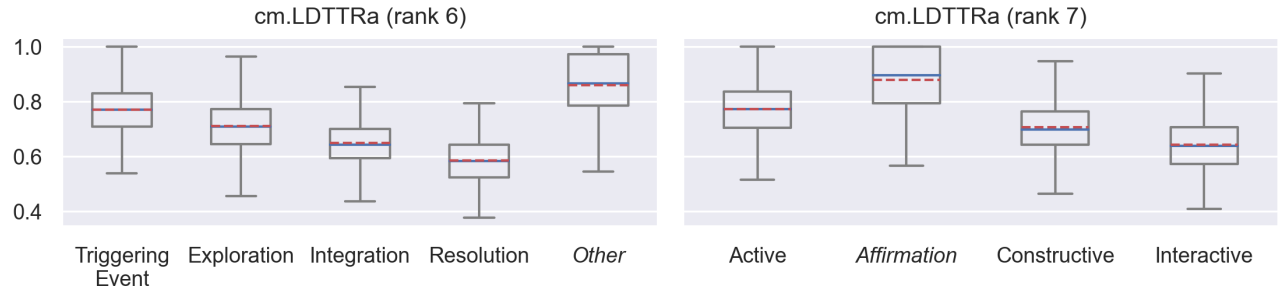


(c) Number of words: Coh-Metrix label `DESWC`

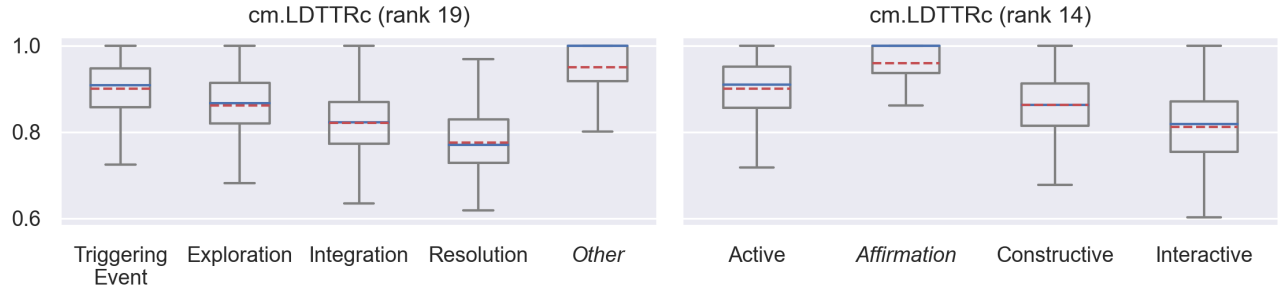
Figure 7. Box plots for the Coh-Metrix features that appear in the top 20 most predictive features in each model in Experiment 1, listed alphabetically by feature name for ease of reference. Plots use the same scale to show the distribution of feature values across (left) the CoI phases of cognitive presence; and (right) the ICAP modes of cognitive engagement. In each plot, the box extends from the lower to upper quartile values of the data, with a solid line at the median. The whiskers extend from the box to show the range of the data. The mean is shown as a broken line. The rank in Experiment 1 is given for reference, and features that are not in the top 20 for one of the frameworks are shown slightly faded.



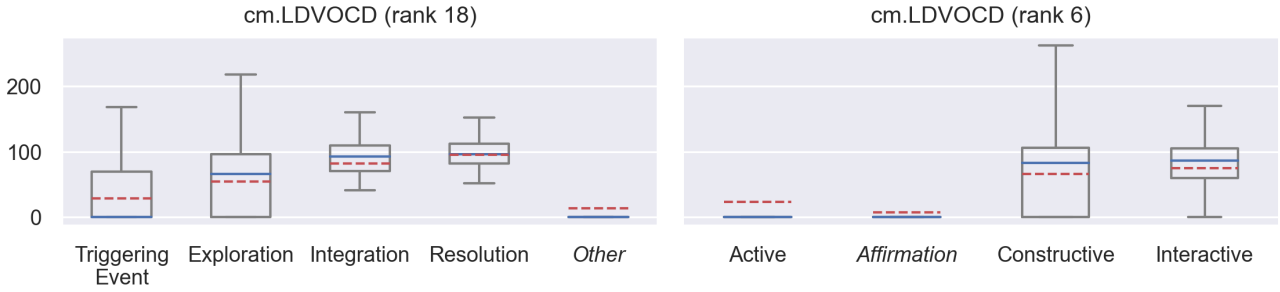
(d) SD of word length in letters: Coh-Metrix label `DESWLltd`



(e) Lexical diversity, all words: Coh-Metrix label `LDTTRa`

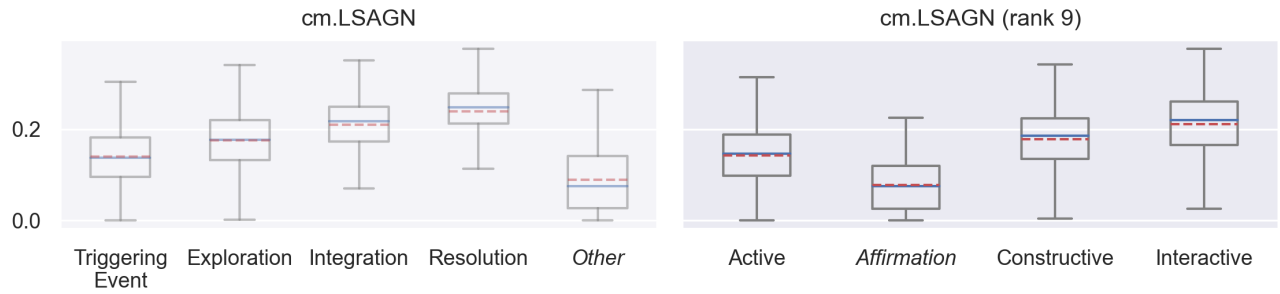


(f) Lexical diversity, content words: Coh-Metrix label `LDTTRc`

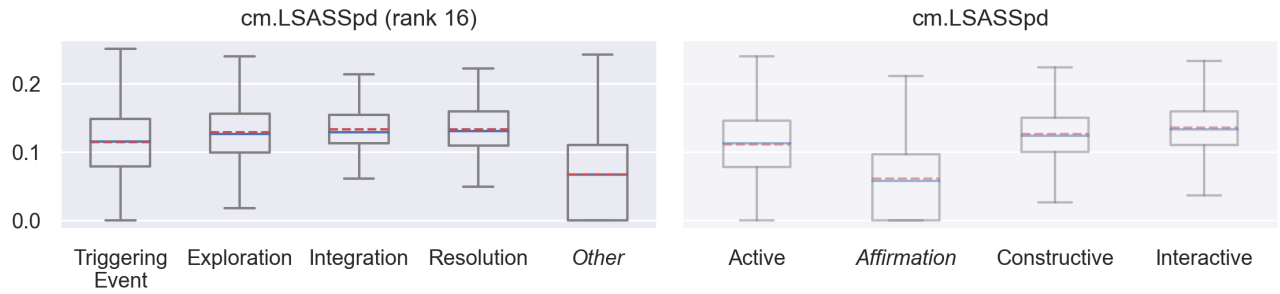


(g) Lexical diversity, VOCD: Coh-Metrix label `LDVOCD`

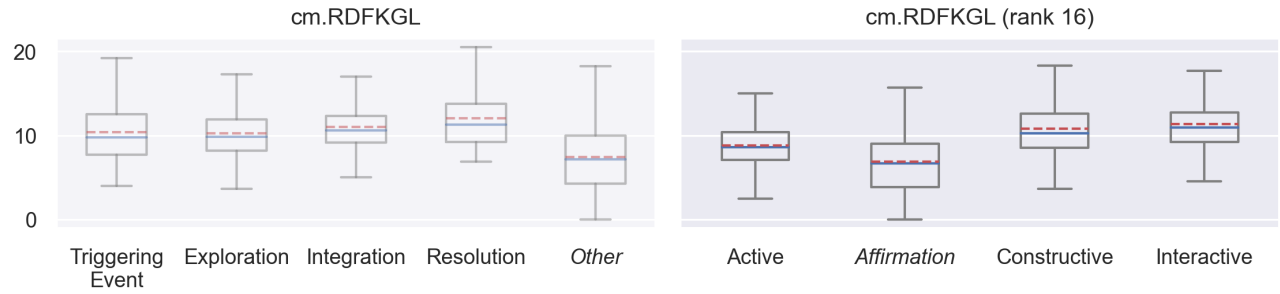
Figure 7. Box plots for the most predictive Coh-Metrix features in Experiment 1, listed alphabetically by feature name, showing their distribution across (left) the CoI phases of cognitive presence, and (right) the ICAP modes of cognitive engagement, on the same scale (cont.)



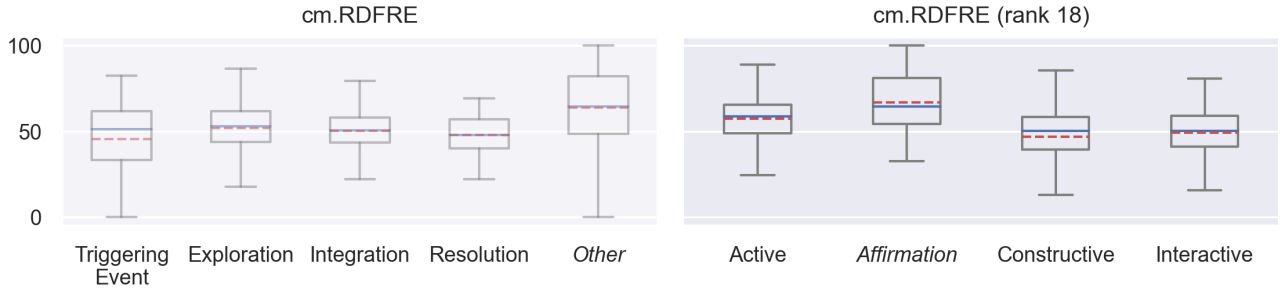
(h) LSA given-new ratio: Coh-Metrix label `LSAGN`



(i) SD of LSA overlap in paragraph: Coh-Metrix label `LSASSpd`

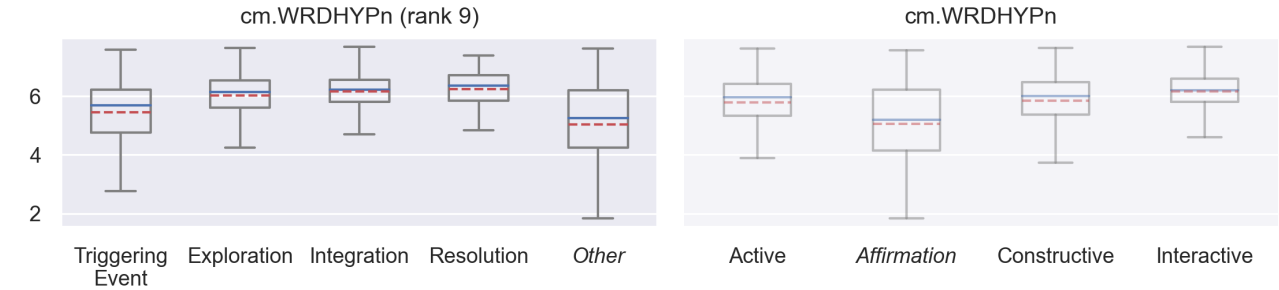


(j) Flesch-Kincaid Grade Level score: Coh-Metrix label `RDFKGL`

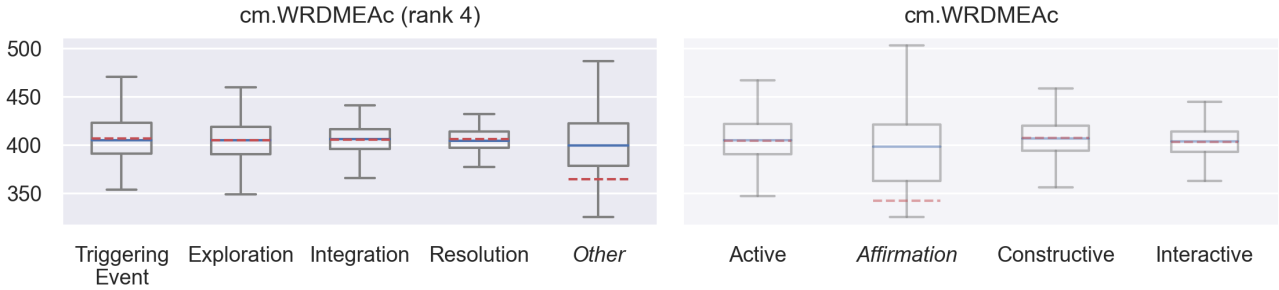


(k) Flesch Reading Ease score: Coh-Metrix label `RDFRE`

Figure 7. Box plots for the most predictive Coh-Metrix features in Experiment 1, listed alphabetically by feature name, showing their distribution across (left) the CoI phases of cognitive presence, and (right) the ICAP modes of cognitive engagement, on the same scale (cont.)

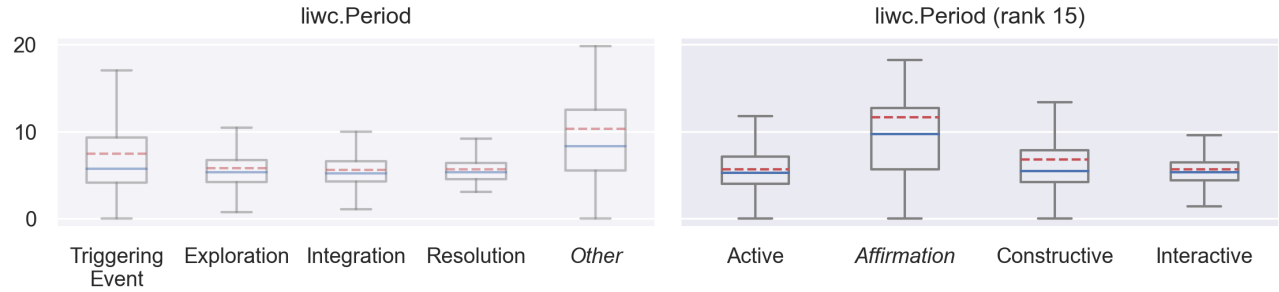


(l) Hypernyms for nouns: Coh-Metrix label `WRDHYPn`



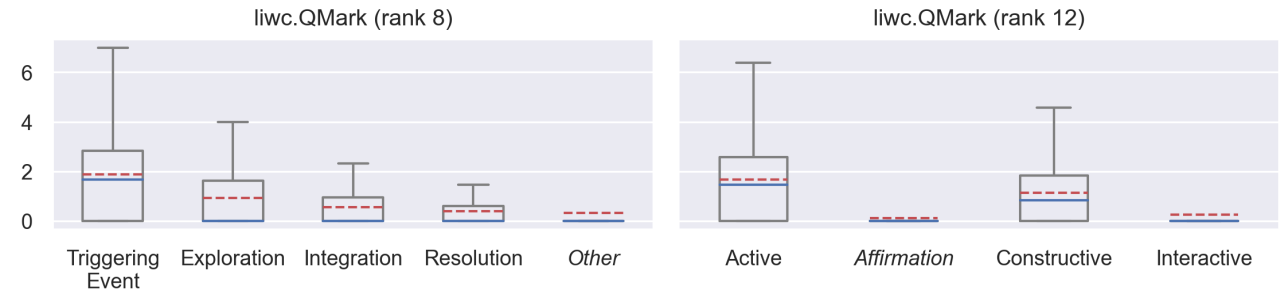
(m) Meaningfulness: Coh-Metrix label `WRDMEAc`

Figure 7. Box plots for the most predictive Coh-Metrix features in Experiment 1, listed alphabetically by feature name, showing their distribution across (left) the CoI phases of cognitive presence, and (right) the ICAP modes of cognitive engagement, on the same scale (cont.)

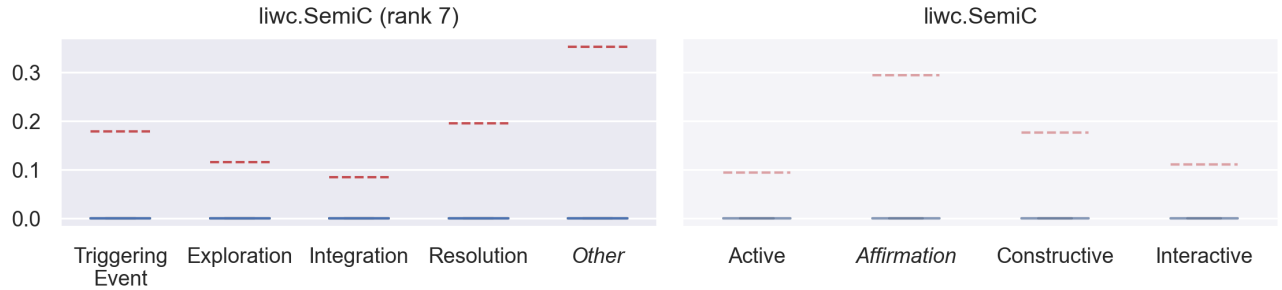


(a) Number of periods: LIWC label `Period`

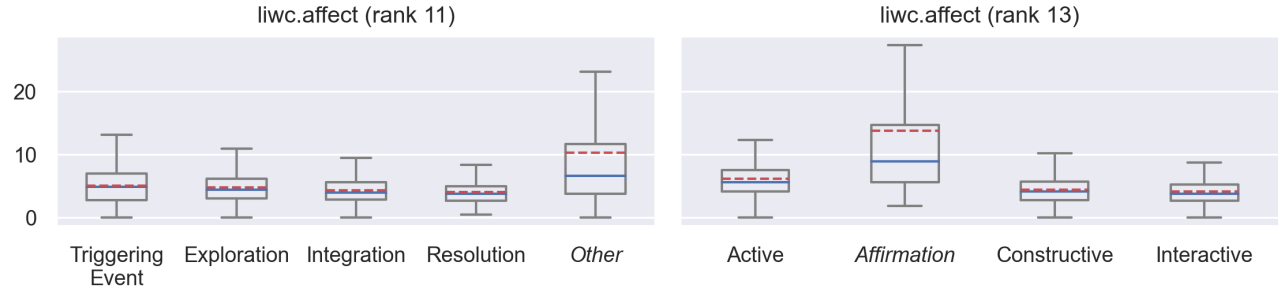
Figure 8. Box plots for the LIWC features that appear in the top 20 most predictive features in each model in Experiment 1, listed alphabetically by feature name for ease of reference. Plots use the same scale to show the distribution of feature values across (left) the CoI phases of cognitive presence; and (right) the ICAP modes of cognitive engagement. In each plot, the box extends from the lower to upper quartile values of the data, with a solid line at the median. The whiskers extend from the box to show the range of the data. The mean is shown as a broken line. The rank in Experiment 1 is given for reference, and features that are not in the top 20 for one of the frameworks are shown slightly faded.



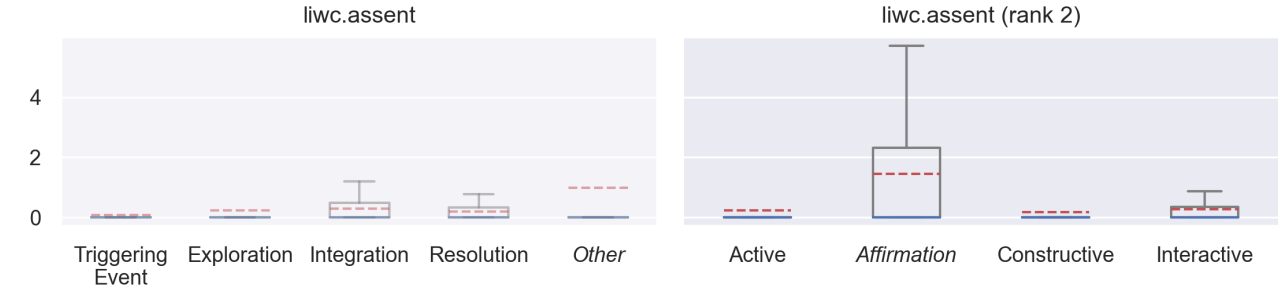
(b) Number of question marks: LIWC label `QMark`



(c) Number of semi-colons: LIWC label `SemiC`

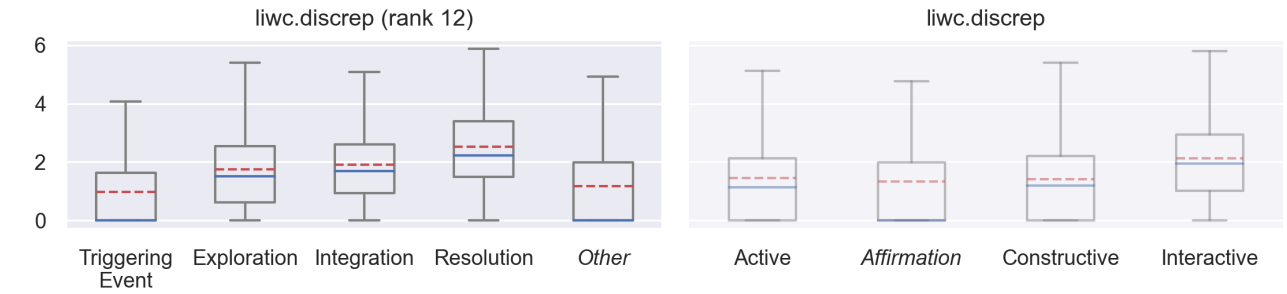


(d) Number of affective process words: LIWC label `affect`

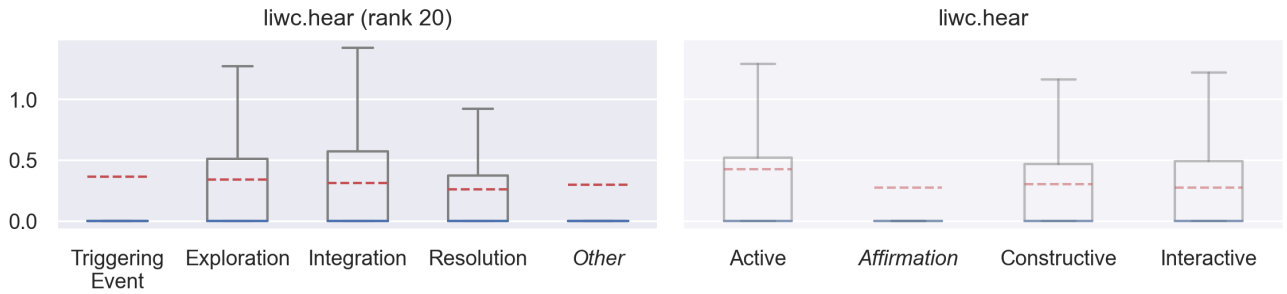


(e) Number of expressions of assent: LIWC label `assent`

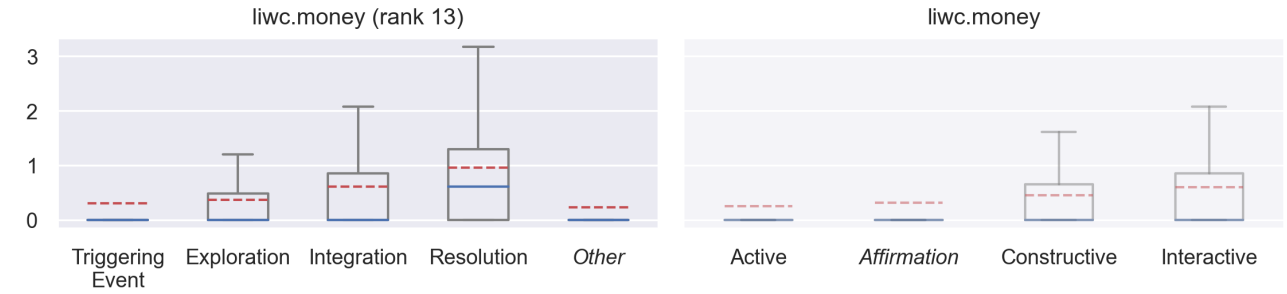
Figure 8. Box plots for the most predictive LIWC features in Experiment 1, listed alphabetically by feature name, showing their distribution across (left) the CoI phases of cognitive presence, and (right) the ICAP modes of cognitive engagement, on the same scale (cont.)



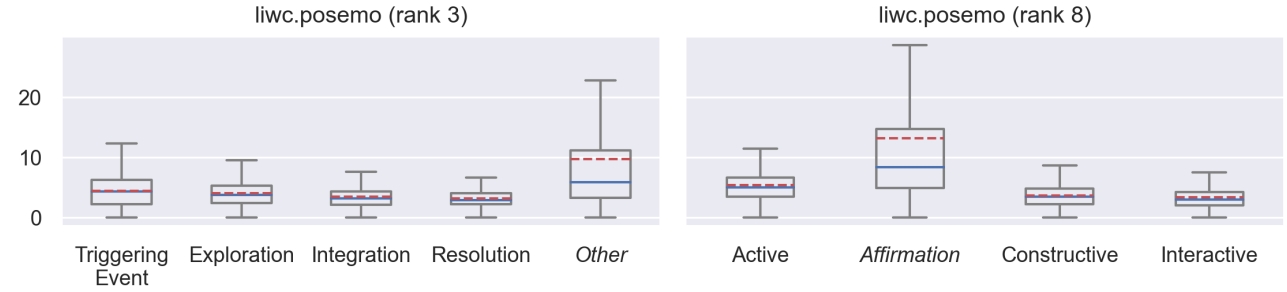
(f) Number of discrepancy words: LIWC label `discrep`



(g) Number of hearing-related words: LIWC label `hear`

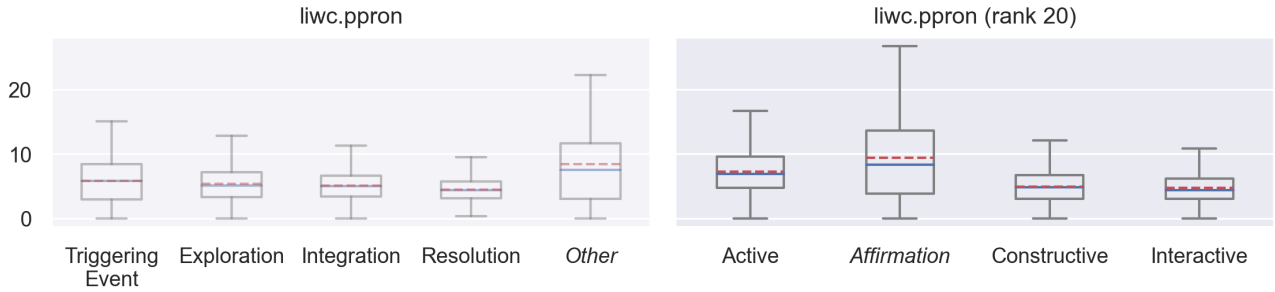


(h) Number of money words: LIWC label `money`

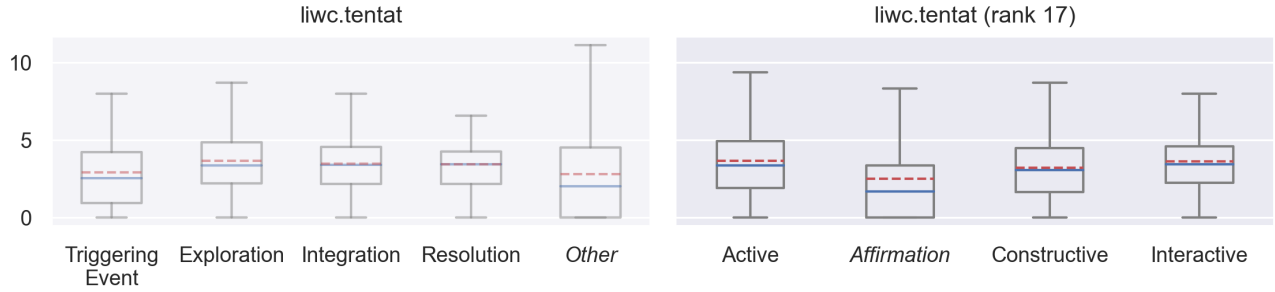


(i) Number of positive emotion words: LIWC label `posemo`

Figure 8. Box plots for the most predictive LIWC features in Experiment 1, listed alphabetically by feature name, showing their distribution across (left) the CoI phases of cognitive presence, and (right) the ICAP modes of cognitive engagement, on the same scale (cont.)

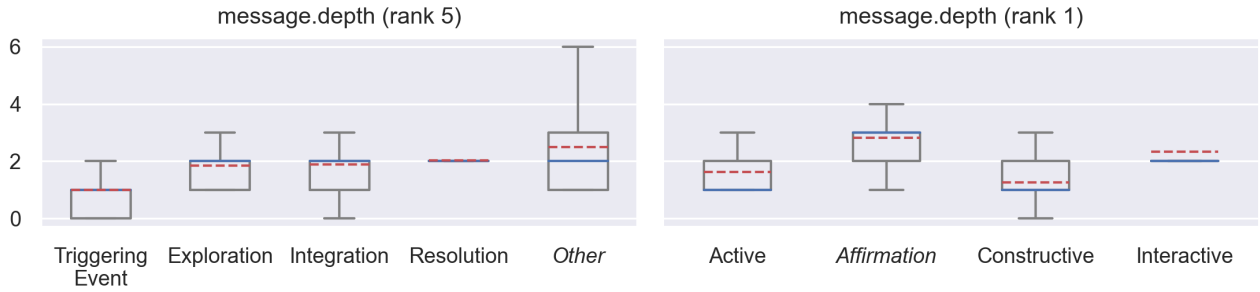


(j) Number of personal pronouns: LIWC label `ppron`



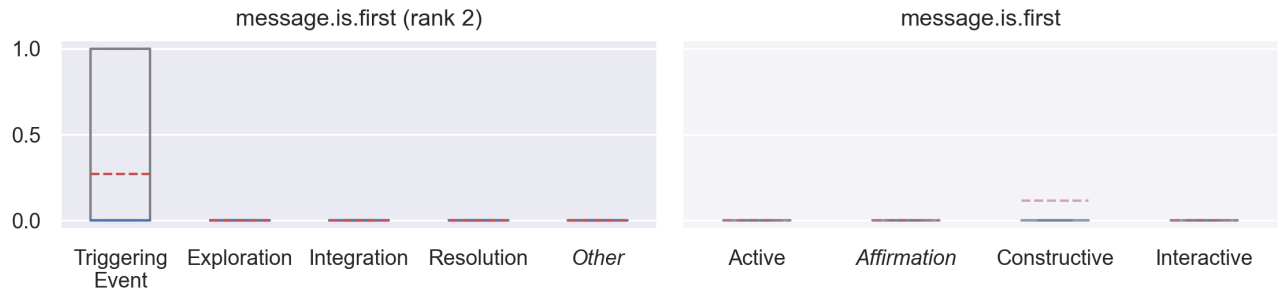
(k) Number of tentative words: LIWC label `tentat`

Figure 8. Box plots for the most predictive LIWC features in Experiment 1, listed alphabetically by feature name, showing their distribution across (left) the CoI phases of cognitive presence, and (right) the ICAP modes of cognitive engagement, on the same scale (cont.)

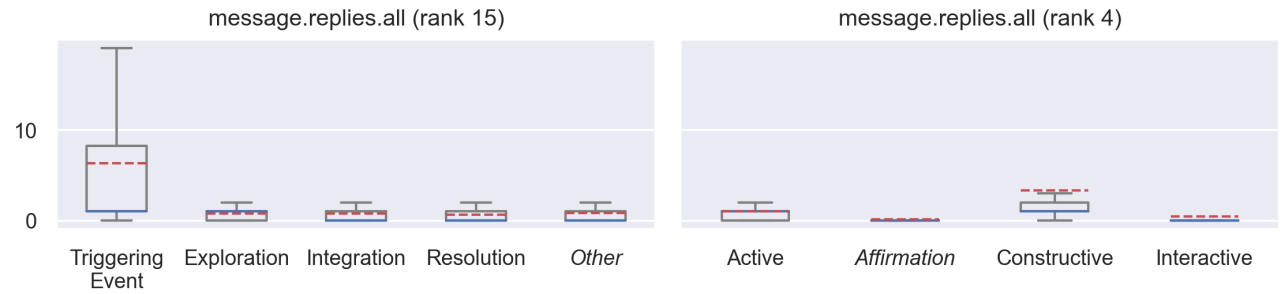


(a) Message depth in discussion: label `message.depth`.

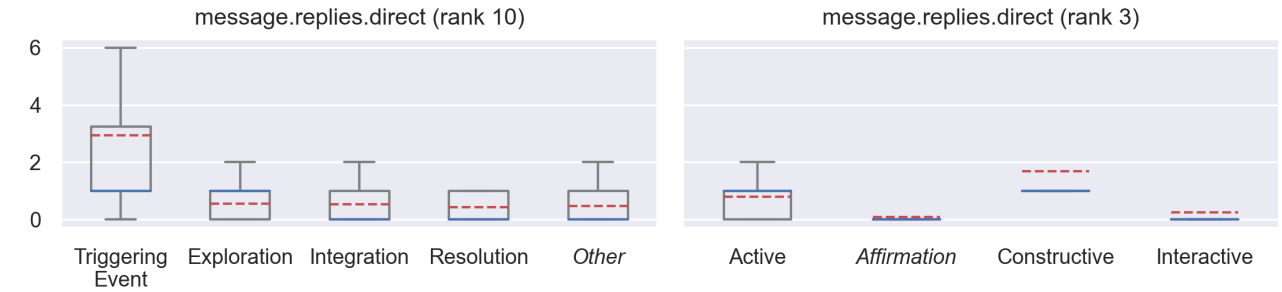
Figure 9. Box plots for the structural features that appear in the top 20 most predictive features in each model in Experiment 1, listed alphabetically by feature name for ease of reference. Plots use the same scale to show the distribution of feature values across (left) the CoI phases of cognitive presence; and (right) the ICAP modes of cognitive engagement. In each plot, the box extends from the lower to upper quartile values of the data, with a solid line at the median. The whiskers extend from the box to show the range of the data. The mean is shown as a broken line. The rank in Experiment 1 is given for reference, and features that are not in the top 20 for one of the frameworks are shown slightly faded.



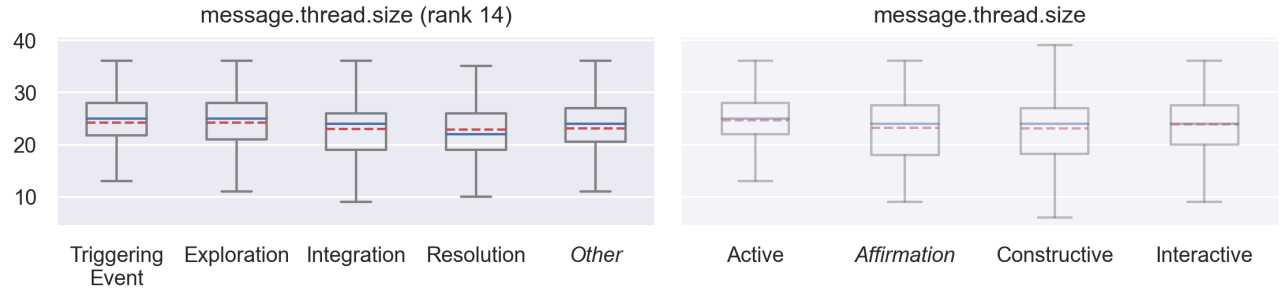
(b) First message: label `message.is.first`



(c) Total number of replies: label `message.replies.all`

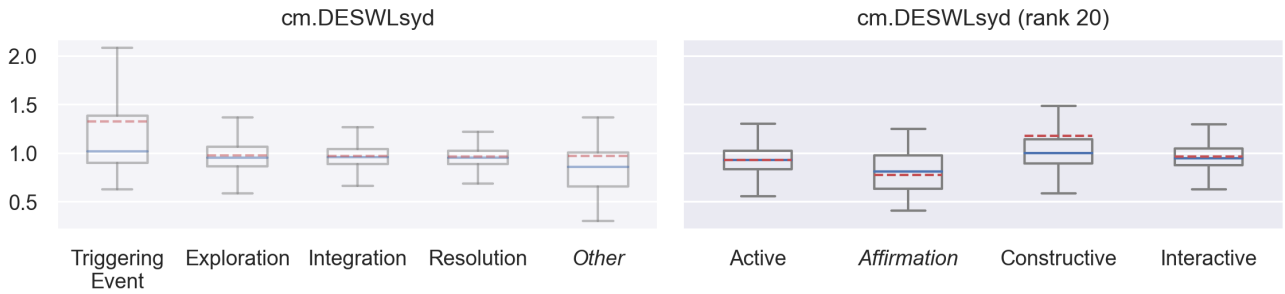


(d) Number of direct replies: label `message.replies.direct`

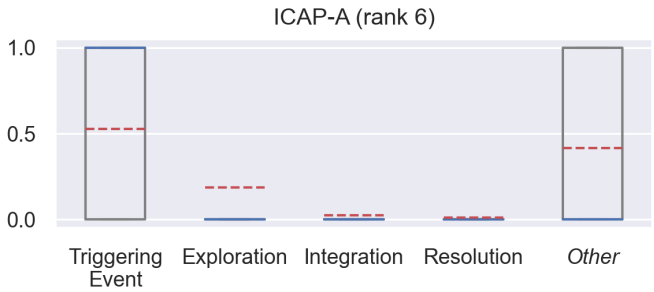


(e) Discussion size: label `message.thread.size`

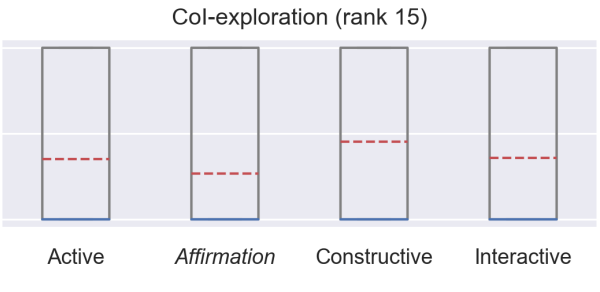
Figure 9. Box plots for the most predictive structural features in Experiment 1, listed alphabetically by feature name, showing their distribution across (left) the CoI phases of cognitive presence, and (right) the ICAP modes of cognitive engagement, on the same scale (cont.)



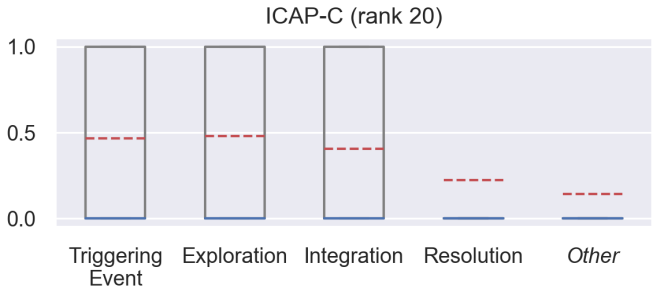
(a) SD of the mean number of syllables: Coh-Metrix label `DESWLsyd`



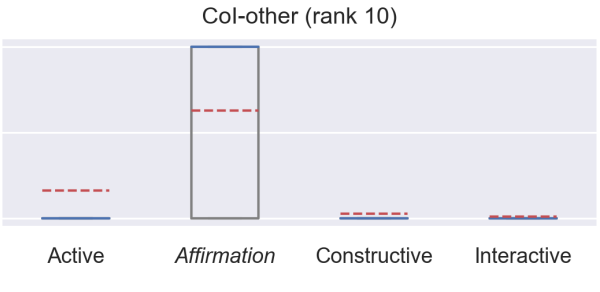
(b) Labelled as *Active*: label `ICAP-A`



(c) Labelled as *Exploration*: label `CoI-exploration`

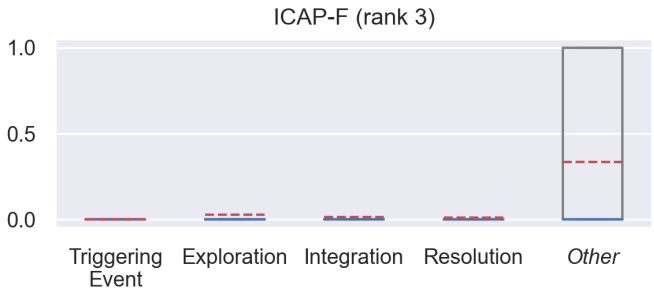


(d) Labelled as *Constructive*: label `ICAP-C`

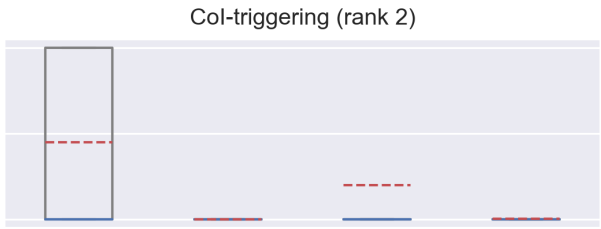


(e) Labelled as *Other*: label `CoI-other`

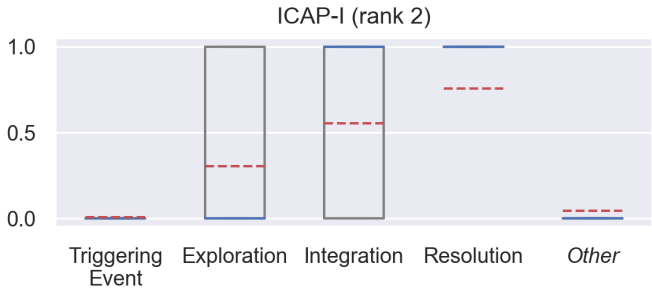
Figure 10. Box plots for the top 20 most predictive features for each framework in Experiment 2, excluding those that were in the top 20 for Experiment 1 (shown in Figures 7, 8, and 9). Plots use the same scale to show the distribution of feature values across (left) the CoI phases of cognitive presence; and (right) the ICAP modes of cognitive engagement. In each plot, the box extends from the lower to upper quartile values of the data, with a solid line at the median. The whiskers extend from the box to show the range of the data. The mean is shown as a broken line. The rank in Experiment 2 is given for reference, and features that are not in the top 20 for one of the frameworks are shown slightly faded.



(f) Labelled as *Affirmation*: label `ICAP-F`



(g) Labelled as *Triggering Event*: label `CoI-triggering`



(h) Labelled as *Interactive*: label `ICAP-I`

Figure 10. Box plots for the top 20 most predictive features in Experiment 2, excluding those that were in the top 20 for Experiment 1, showing their distribution across (left) the CoI phases of cognitive presence, and (right) the ICAP modes of cognitive engagement, on the same scale (cont.)

Table 21. Features used and metrics reported in Experiment 1 and in previous studies using random forest classifiers to label the CoI phases of cognitive presence.

	Experiment 1	Kovanović et al. (2016)	Farrow et al. (2019)	Neto et al. (2018)	Barbosa et al. (2020)
Features used					
Coh-Metrix	106	106	106	48	38
LIWC	91	91	91	24	64
Message depth	✓	✓	✓	✓	✓
First/ last message	✓	✓	✓	✓	–
Number of replies	direct + indirect	✓	✓	✓	✓
Cosine similarity to previous/ next messages	–	✓	✓	✓	✓
Count of named entities	–	relevant	relevant	all	all
Internal coherence	–	LSA	LSA	embeddings	LSA
Metrics reported					
Cohen's κ	0.358	0.63	0.38	0.72	0.53
Macro-averaged F_1	0.515	–	0.54	0.63	–

Table 22. The top 20 features from Experiment 1 shown with their ranks in previous studies where random forest classifiers were used to label the CoI phases of cognitive presence. The label – indicates that the feature was ranked outside the top 20, while the label × indicates that the feature was not used in the model, to the best of our understanding. The top ranking features from Farrow et al. (2019) were not previously published.

#	Feature	Description	Kovanović et al. (2016)	Farrow et al. (2019)	Neto et al. (2018)	Barbosa et al. (2020)
1	`cm.DESWC`	Number of words	1	2	10	1
2	`message.is.first`	First message	–	3	–	×
3	`liwc.posemo`	Number of +ve emotion words	–	9	×	×
4	`cm.WRDMEAc`	Meaningfulness	–	12	×	×
5	`message.depth`	Message depth in discussion	4	4	3	3
6	`cm.LDTTRa`	Lexical diversity, all words	3	6	16	–
7	`liwc.SemiC`	Number of semi-colons	–	5	–	–
8	`liwc.QMark`	Number of question marks	7	10	1	4
9	`cm.WRDHYPn`	Hypernyms for nouns	–	8	–	–
10	`message.replies.direct`	Number of direct replies	×	×	×	×
11	`liwc.affect`	Number of affective process words	–	15	×	×
12	`liwc.discrep`	Number of discrepancy words	–	7	×	×
13	`liwc.money`	Number of money words	10	11	×	×
14	`message.thread.size`	Discussion size	×	×	×	×
15	`message.replies.all`	Total number of replies	13	16	–	–
16	`cm.LSASSpd`	SD of LSA overlap in paragraph	–	19	–	–
17	`cm.DESWLltd`	SD of word length in letters	–	13	–	–
18	`cm.LDVOCD`	Lexical diversity, VOCD	9	18	14	–
19	`cm.LDTTRc`	Lexical diversity, content words	5	–	15	–
20	`liwc.hear`	Number of hearing-related words	–	–	×	×
Features from Experiment 1 ranked in the top 20			8/18	16/18	6/12	3/11