This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Modelling the Cognitive Quality of Student Contributions to Online Discussion Forums

Elaine Margaret Farrow

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2023
Abstract

Understanding how students can develop their critical thinking skills and engage in social knowledge construction through discussion with their peers is important for both educators and researchers. As asynchronous online discussion forums become increasingly common across educational settings of all kinds, there is a growing need to identify the characteristics of effective discussions that are associated with learning gains. Such findings can inform the way discussion-based assignments are framed and assessed and can provide evidence about the efficacy of instructional interventions. While many messages are purely social in nature, others demonstrate intellectual engagement with the subject matter of the course, to a greater or lesser extent – the cognitive quality of the message. However, it is not straightforward to measure cognitive quality. Previous research has defined cognitive engagement based on the visible learning behaviours of students; and identified distinct phases of cognitive presence commonly seen in collaborative online discussions among groups of participants. Little prior work has brought together insights from both individual learning behaviours and group discussion dynamics, a gap this thesis aims to fill.

This thesis introduces a two-dimensional measure of cognitive quality, making use of constructs from two well-supported educational frameworks: the Interactive-Constructive-Active-Passive framework and the Community of Inquiry framework. Using a pseudonymised set of messages that were labelled using both frameworks, the thesis explores how attributes of the dialogue were correlated with cognitive quality. Message quality was found to depend more on the nested discussion structure than on chronological order. As previously seen with other frameworks, the same messages tended to be identified as high-quality by both frameworks, while there was more variation among mid- and lower-quality messages. The thesis goes on to investigate the potential moderating effects of two instructional interventions: assigning roles to students within the asynchronous online discussions; and an external facilitation intervention, introducing guidelines that aimed to enhance the quality of students’ self-regulation. Using a novel network analytic approach, the external facilitation was observed to moderate the associations between the frameworks, while no such change was seen with the role assignment. Finally, the thesis finds that the order in which students took on the assigned roles had minimal impact on the cognitive quality of their contributions to the discussion. This thesis contributes new, actionable findings about the factors that influence the cognitive quality of student contributions to asynchronous online discussions and concludes with a discussion of future research directions.
Lay Summary

Discussion forums are common in many educational settings. They allow students to learn from each other and to develop their understanding by challenging ideas and building on each other’s contributions. Students who engage with the forums tend to learn more. However, it is not obvious how to assess the quality of a forum message, or to measure the value it adds to the discussion. Identifying what makes a message high quality is the focus of this thesis and is useful for both educators and learners.

Earlier studies proposed many different ways of labelling messages. For example, the ICAP framework labelled each message based on what the student was doing – copying or paraphrasing course materials; working alone to create something new, like a summary; or challenging or extending the ideas of a partner. The Community of Inquiry framework looked, instead, at the way discussions often progress through distinct phases – asking questions, exploring new ideas, making connections between ideas, then proposing possible solutions – and labelled each message based on the relevant phase. Labels like these can be used as indicators of message quality.

This thesis used labels from the ICAP and Community of Inquiry frameworks together to create a two-dimensional quality score that captured both the student’s own engagement and their contribution to the progress of the overall discussion. A set of messages collected from a real discussion forum was given both kinds of labels. We found that the labels that indicated the highest quality in each framework were generally given to the same messages, while the mid-level and lower-level labels varied more. Initially, messages where the students displayed reasoning tended to be limited to exploring a single idea. After changing the assignment instructions to encourage all students to write higher quality messages, reasoning was more often used to link the new idea with other content from the course. Students who took on the role of the expert when discussing a particular topic wrote higher quality messages about that topic than the non-experts, no matter which role they experienced first.

Since there are many different ways to express the same idea, this thesis looked beyond the individual words in a message to consider how other attributes – such as word categories, message length, and timing – were associated with quality. It found that messages that were more deeply nested in discussion threads were generally higher quality. Messages that were sent later in time varied in quality. Students should therefore be encouraged to go deeper and develop existing arguments further. This will improve the overall quality of the discussion more than asking new top-level questions.
This thesis is dedicated to my family,
with all my love
I want to start by expressing my heartfelt thanks to my two supervisors, Professor Johanna Moore and Professor Dragan Gašević, who guided and encouraged me, always sharing their wisdom and experience generously. I have learned so much from them both. They made time to connect with me regularly, even when Covid restrictions prevented us from meeting together physically.

My thanks and appreciation also go to Professor Henry Thompson, who stepped in to provide local support and a listening ear while both my supervisors were on the other side of the world.

I am grateful for all my wonderful friends and colleagues, some of whom deserve particular thanks for reading drafts of my work and giving helpful and constructive feedback: Amy Isard, Vashti Galpin, and Yi-Shan Tsai. Other colleagues and fellow-students became my conference buddies and hackathon team-mates, practising our presentations together and sharing ideas about learning analytics and ENA – thank you all, John Saint, Yizhou Fan, Ana Hibert, Alexia Revueltas Roux, Sarah Austin, Diego Rates, and Barbara Dzieciatko.

Other groups of people who played an important role in helping me through this journey include my classmates in the Data Science CDT, who made the PhD experience less lonely, and my Zoom pub pals, who kept me grounded and reminded me that there’s life beyond studying.

Most of all, I want to thank my family for bearing with me – especially Edmund Farrow, who did not expect to end up sharing a house with four students.

**Funding acknowledgement**: The work reported in this thesis was supported 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.
Declaration of Authorship

I declare that this thesis was composed by myself and that the work contained in it has not been submitted, in whole or in part, for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contributions and those of the other authors have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

This thesis includes verbatim copies of four peer-reviewed publications, all produced under joint authorship with my supervisors and reproduced here with permission of the rights holders:


I declare that I contributed substantially to all four publications (i.e., over 50% of the work done) and was involved in all phases of the research process. The ideas and analysis in each paper were developed and discussed with my supervisors. The original idea, the experiments, and the writing were my own work.

Elaine Farrow
# Table of Contents

Abstract iii  
Lay summary iv  
Acknowledgements vi  
Declaration of authorship viii  

1 Introduction 1  
1.1 Research questions 11  
1.2 Methodology 12  
1.3 Thesis structure and overview 19  

2 Data preparation 25  
2.1 Introduction 25  
2.2 Peer-reviewed publication: Names, Nicknames, and Spelling Errors: Protecting Participant Identity in Learning Analytics of Online Discussions 35  
2.3 Summary of contributions 48  

3 Markers of cognitive quality 49  
3.1 Introduction 49  
3.2 Peer-reviewed publication: Markers of Cognitive Quality in Student Contributions to Online Course Discussion Forums 51  
3.3 Summary of contributions 80  

4 Integrating multiple quality measures 83  
4.1 Introduction 83  
4.2 Peer-reviewed publication: A Network Analytic Approach to Integrating Multiple Quality Measures for Asynchronous Online Discussions 86  
4.3 Summary of contributions 99
# List of Figures

1.1 The CoI model, reprinted from Garrison et al. (1999) .......................... 4  
1.2 The four phases of cognitive presence within the CoI model ................. 4  
1.3 The hierarchical modes of cognitive engagement in the ICAP framework 5  

2.1 Patterns of interaction between participants ........................................ 28  
2.2 The extended ICAP annotation flow chart, reprinted from Yoge v et al.  
(2018) ........................................................................................................ 32  
2.3 The extended cognitive engagement annotation flow chart developed  
and used in this thesis ............................................................................... 33  

4.1 Higher-order and lower-order thinking across six course sessions ....... 85  
4.2 Subtracted networks for each intervention ........................................... 101  

B.1 Update of Figure 4 from Farrow et al. (2022) ................................. 139
# List of Tables

1.1 Overview of the research questions addressed in each chapter . . . . . 19

2.1 Cognitive presence descriptors and indicators, reprinted from Garrison (2011) ................................. 30

2.2 The extended cognitive engagement taxonomy used in this thesis . . . 31

3.1 Dialogue attributes found to be most predictive ....................... 81

4.1 Indicators of higher-order and lower-order thinking .................... 84

5.1 Single duty roles used in previous studies .............................. 104

5.2 Mapping from single-duty roles to the composite roles used in this thesis 105
Chapter 1

Introduction

Technology increasingly plays a central role in the educational experience for many students, and its importance is widely recognised by researchers. The significance of online discussion forums, where students can interact with one another and with their instructors, is of particular note (Garrison et al., 1999; Garrison, 2011; Wang et al., 2015; Wise & Cui, 2018). Such forums are used in a wide variety of educational settings (Rosé et al., 2017), from fully-online distance-learning courses (Gašević, Adesope, et al., 2015; Neto et al., 2018) and massive open online courses (MOOCs) (Wang et al., 2016b) to traditional on-campus classes (Ryan, 2013). Forums allow students to interact productively with peers and instructors, irrespective of physical location or time zone. While many discussion threads focus on administrative matters or are purely social in nature (Cui et al., 2017), it is also common for constructive participation in a forum to be listed as a course requirement and to receive course credit (Gilbert & Dabbagh, 2005; Gašević, Adesope, et al., 2015; Neto et al., 2018; Penteado et al., 2019). In such cases, the discussion topic may be provided in advance (Gilbert & Dabbagh, 2005). Students may also receive guidance about the nature and expected quantity of their contributions (Gilbert & Dabbagh, 2005; Gašević, Adesope, et al., 2015). For example, each student might be instructed to start one new message thread and to reply to three messages written by their peers.

In addition to their primary role in supporting education through interaction and collaboration, the messages exchanged in discussion forums can also be used to inform research (Meyer, 2004; De Wever et al., 2006; Rosé et al., 2008; Knight et al., 2014; Gašević et al., 2017; Graesser et al., 2017; Rosé et al., 2017; Trausan-Matu & Slotta, 2021). Discussion forum data is especially valuable for investigating how people work together to construct knowledge and gain greater understanding of
course topics, building on each other’s contributions (Garrison et al., 2001; Meyer, 2004; Schrire, 2004; Gilbert & Dabbagh, 2005; O’Donnell & Hmelo-Silver, 2013; Suthers et al., 2013; Schecter & Contractor, 2017; Kimmerle et al., 2021; Stahl & Hakkarainen, 2021). Many frameworks have been developed to allow researchers to analyse student interactions in a systematic way (Garrison et al., 1999; Ferguson et al., 2013; Chi & Wylie, 2014; Cui & Wise, 2015; Wise et al., 2016; Cross et al., 2017; Harrak et al., 2018; Hu et al., 2020). These frameworks provide coding schemes that can be used to assign labels to the data at different levels of granularity (De Wever et al., 2006) – for example, messages, sentences, or conversation threads – based on the written content and the context. In some studies, messages were split into smaller segments, each of which was labelled, so that every message had an ordered sequence of labels (Corich et al., 2006; Rosé et al., 2008). The most commonly chosen unit of analysis is the individual message (Garrison et al., 2001; Schellens et al., 2007; Hu, Donald, & Giacaman, 2021). The framework labels can be used to identify conversation threads that are developing appropriately over time and those that have stalled or are off-task (Hosmer & Lee, 2021).

Manual content analysis techniques are commonly used to assign labels to a discussion, using indicators defined by the relevant framework (Schrire, 2004; Hu et al., 2020; Hu, Donald, & Giacaman, 2021). Self-reported measures based on validated survey instruments can also be used (Arbaugh et al., 2008). However, manual content analysis is slow and expensive (Rosé et al., 2008) and self-reports are impractical for real-time monitoring, meaning that analysis using the labels is only possible after a course has ended. Aiming to alleviate the labelling bottleneck, supervised classification methods have recently shown promise in being able to label new data correctly based on a corpus of labelled examples (Rosé et al., 2008; Ferguson et al., 2013; Waters et al., 2015; Kovanović et al., 2016; Wise et al., 2016; Neto et al., 2018; Yogev et al., 2018; Hayati et al., 2020; Hu, Ferreira Mello, et al., 2021). Trained classifiers, such as random forest models (Breiman, 2001), use labelled examples to learn associations between attributes of the dialogue and the framework labels. Attributes that are used as model features can relate to the structure of the dialogue or of the individual message, such as message length, or they can represent linguistic patterns within the text (Trausan-Matu & Slotta, 2021) that identify relevant discourse actions (Rosé et al., 2008). Once trained, the classifiers can be used to label previously unseen messages. As the number of discussion messages gets larger and it becomes increasingly difficult for human tutors to monitor forums effectively (Cui & Wise, 2015; Lee et al., 2022),
automatically assigned labels could help instructors to see where their input is most needed (Rosé et al., 2008; Cui & Wise, 2015; Harrak et al., 2018; Ludvigsen et al., 2021; Chen & Teasley, 2022). Classifiers that were deployed during the course could also be used to trigger automated support for learning (Rosé et al., 2017). A secondary benefit of trained random forest classifiers is that they can provide further insights into patterns within the data by revealing which of the dialogue attributes are most discriminative, thereby providing new insights into the framework constructs themselves (Rosé et al., 2008; Kovanović et al., 2016; Neto et al., 2018; Hu et al., 2022b).

While some messages are purely social in nature, or focus on practical and administrative matters (Cui et al., 2017), others demonstrate intellectual engagement with the subject matter of the course, to a greater or lesser extent (Garrison et al., 2001; Gilbert & Dabbagh, 2005; Cui et al., 2017; Hayati et al., 2020). Although frameworks differ, a message that introduces new content or novel ideas to the discussion will typically be rated more highly than another message that merely restates or paraphrases previously given information without adding anything new (Bloom, 1956; Gilbert & Dabbagh, 2005; Schrire, 2006; Yoge et al., 2018; Kimmerle et al., 2021; Chen & Teasley, 2022; Scardamalia & Bereiter, 2022). This thesis refers to the demonstration of higher-order and critical thinking skills through discussion as **cognitive quality**, defined in more detail in Chapter 3.

This thesis makes use of two popular and widely-used frameworks for classifying the cognitive quality of student contributions to asynchronous online discussions. The first of these is the **Community of Inquiry (CoI) framework** (Garrison et al., 1999), illustrated in Figure 1.1, where three related *presences* function together to create a well-balanced educational experience. Of these three, *cognitive presence* is considered to be the most fundamental for learning (Garrison et al., 1999), defined as “the extent to which the participants in any particular configuration of a community of inquiry are able to construct meaning through sustained communication” (Garrison et al., 1999, p. 89). CoI is one of the most well-studied theoretical frameworks in online education (Gašević, Adesope, et al., 2015; R. Ferreira et al., 2018) and cognitive presence has been widely used to assess higher-order thinking skills in studies using both manual content analysis (Garrison et al., 2001; Corich et al., 2006; Arbaugh et al., 2008) and trained machine learning classifiers (T. E. McKlin, 2004; Waters et al., 2015; Kovanović et al., 2016; Lee et al., 2022).

According to CoI, in the idealised case, group discussions progress sequentially through the four *phases of cognitive presence*, which build on one another (Figure 1.2)
Figure 1.1: The CoI model, reprinted from Garrison et al. (1999), illustrating the contributions of social presence, cognitive presence, and teaching presence to the overall educational experience. Reprinted from The Internet and Higher Education, 2(2–3), Garrison, D. R., Anderson, T., & Archer, W., Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education, 87–105, 1999, with permission from Elsevier.

Figure 1.2: The four phases of cognitive presence within the CoI model. The cycle begins with a triggering event. A successful resolution may trigger a new cycle of inquiry (Garrison, 2011).
Figure 1.3: The hierarchical modes of cognitive engagement in the ICAP framework.

– although there is a tendency for educators “to do the first two phases very well, the third phase less well, and the last phase hardly at all.” (Garrison et al., 2001, p. 53):
  • **Triggering Event**: presenting information about an issue and inviting questions.
  • **Exploration**: proposing ideas, sharing personal experience, and brainstorming possible solutions and explanations.
  • **Integration**: connecting ideas together and developing them further to construct meaningful solutions and explanations.
  • **Resolution**: reaching consensus and critically assessing a proposed solution.

The second framework that is used in this thesis to define and measure cognitive quality is the **Interactive-Constructive-Active-Passive (ICAP) framework** (Chi & Wylie, 2014), illustrated in Figure 1.3. The ICAP framework was developed through a process of meta-analysis and empirical validation, grounded in classroom studies and lab experiments. Although ICAP was not originally designed with online learning in mind, it has been adapted and used as a foundation for studies on computer-supported collaborative learning (CSCL), where it can be applied to learning activities as a whole, or to differentiate the behaviour of individual learners within a group (Wang et al., 2015). Some studies have applied the ICAP framework to the individual messages within online discussions (Wang et al., 2016b; Taskin et al., 2019). As with CoI, studies using ICAP have used both manual content analysis (Wang et al., 2016b; Vellukunnel et al., 2017) and automated labelling (Wang et al., 2015; Yogev et al., 2018; Taskin et al., 2019; Hayati et al., 2020).

ICAP defines four **modes of cognitive engagement**, which build on one another in a hierarchical manner:
  • **Passive**: reading or watching without actively doing anything else.
• **Active**: engaging in some kind of physical action that is related to the course and requires focused attention.

• **Constructive**: generating novel output that is related to the course, going beyond what was already given.

• **Interactive**: extending or challenging the **Constructive** ideas of another participant.

The CoI and ICAP frameworks were selected for use in this thesis because of their proven relevance to asynchronous online discussion forums, their focus on the cognitive aspects of learning, and evidence from previous studies that their constructs can be labelled successfully using machine learning classifiers. Both frameworks have been used extensively in earlier studies, but little prior work has been done to compare them either conceptually or through experimentation. One recent study (Hayati et al., 2020) trained a classifier to categorise individual learners according to the ICAP modes of cognitive engagement and included as model features the number of messages in each of the CoI phases of cognitive presence, but did not comment on any possible alignment between the framework constructs. Another study (Wang et al., 2015), where ICAP was used to explore and analyse student learning behaviour in a MOOC, acknowledged the importance of social presence to support cognitive presence but did not explore the CoI presences any further.

In terms of theoretical orientation and positioning, this thesis sits within the discipline of **learning analytics (LA)** and addresses “data and learning and the complexities that follow from the combination of the two” (Lang et al., 2022, p. 8). LA has been described as an interdisciplinary, “bricolage” field, incorporating theory, design, and data science; all three dimensions are necessary, but the emphasis placed on each will vary according to the focus of the study (Gašević et al., 2017). LA is concerned with both theoretical advances and practical applications; it has “a dual identity as an academic discipline and an endeavour relevant to real design and educational practice.” (Buckingham Shum et al., 2019, p. 2). LA is primarily focused on higher education and similar formal learning settings (Chen & Teasley, 2022) and encompasses large-scale environments including massively open online courses (MOOCs).

When LA is used in conjunction with collaborative learning, the work can be described as **analytics of collaborative learning (ACL)** or as **collaborative learning analytics (CLA)** (Wise et al., 2021). ACL involves the use of computational methods and analytics to improve understanding of collaborative processes, while CLA uses data analytics and LA in real time to support and facilitate the collaboration itself (Wise et al., 2021; Chen & Teasley, 2022). An example of ACL is the use of LA to identify attributes
of the data that are correlated with learning constructs defined by a relevant theoretical framework (Wise et al., 2021); for example, the various studies mentioned earlier in this chapter where machine learning classifiers were used to identify the CoI phases of cognitive presence and the ICAP modes of cognitive engagement in asynchronous online discussions. Another example of recent ACL work is the use of eye-tracking equipment to measure and track the visual attention of pairs of participants as they collaborate, to identify instances of joint attention, and to combine that information with metrics representing semantic features of their conversation (Schneider & Pea, 2015; Olsen et al., 2017; Wise et al., 2021). The findings of studies using ACL may be tested experimentally, by redesigning aspects of a collaborative system and observing whether learning is improved, in a process referred to as “closing the loop” (Koedinger et al., 2013; Koedinger & McLaughlin, 2016). For example, findings regarding visual attention could be tested by manipulating a different collaborative learning situation to promote or hinder the ability of participants to perceive one another’s gaze (Schneider & Pea, 2015). However, unless such system changes happen dynamically and repeatedly, such studies are still classified as ACL rather than CLA. The present work also falls into the ACL category: the collaborative learning activity took place before the analysis reported in this thesis was conceived, and the analysis was carried out offline.

The second category (CLA) involves embedding LA within a collaborative interface so that results can be acted on in real time, perhaps by presenting a visualisation of relevant metrics to participants as a prompt for reflection, or by dynamically adapting system behaviour (Rosé & Dimitriadis, 2021). An important example, demonstrating how CLA can be used to support collaboration within groups of human users, is smart learning spaces, such as Knowledge Forum (Scardamalia & Bereiter, 2005, 2021, 2022). Knowledge Forum is a widely-studied platform aimed at young students (Stahl & Hakkarainen, 2021) that contains a suite of embedded analytics tools for the use of students and teachers (O’Donnell & Hmelo-Silver, 2013; Scardamalia & Bereiter, 2021). In general, smart learning spaces are designed to support collaboration among groups of (human) users (Chen et al., 2021); they can orchestrate the collaboration according to a pre-defined script or can simply provide access to tools, to be used at the discretion of the participants (Chen & Teasley, 2022; Scardamalia & Bereiter, 2022). The Knowledge Forum platform is part of a wider program of Knowledge Building, which originated in the 1980s and emphasises the value of community knowledge creation and the need for students to be given agency to direct their own learning (Scardamalia & Bereiter, 2021).
Chapter 1. Introduction

In comparison with the iterative experimental testing described above for ACL systems, LA tools embedded inside a CLA setup can create a far tighter feedback loop by providing information to participants and to the system itself in real time, potentially triggering multiple configuration changes within the same interaction (Wise et al., 2021). An example would be a system that continuously analyses collaborative interactions within a group of participants and adapts its behaviour to adjust the type and amount of support provided. As described earlier in this chapter, promising work has already been done towards automating the labelling of student messages in asynchronous online discussion forums with indicators of cognitive quality. If such automated labelling became sufficiently reliable, its outputs could be used within a CLA system to provide real-time tailored guidance to participants about how to improve the quality of their contributions, as they move from being “novices” towards becoming “experts” (Knight et al., 2014).

Looking beyond the boundaries of LA to the study of collaborative learning more broadly, other CSCL studies are also relevant as background to this thesis. CSCL predates LA (Cress et al., 2021) and contributed to its formation (Chen & Teasley, 2022). CSCL is itself interdisciplinary and encompasses a multiplicity of concepts, methodologies, and theoretical frameworks (Suthers et al., 2013; Cress et al., 2021; Hmelo-Silver & Jeong, 2021). The breadth of CSCL is viewed as a strength (Ludvigsen et al., 2021); for example in the “productive multi-vocality” project (Suthers et al., 2013), which brought together multiple research teams from different traditions within CSCL to work on the same data sets. Within CSCL, both cognitive and non-cognitive factors are acknowledged as being important for learning through collaboration, and assessment of collaboration is typically conducted across cognitive, social, and motivational dimensions (Rosé et al., 2017). Collaborative learning with peers in a classroom setting often provides social benefits, which can, in turn, improve learning outcomes (O’Donnell & Hmelo-Silver, 2013); conversely, working as a group can sometimes be challenging and damage learners’ motivation (O’Donnell & Hmelo-Silver, 2013). Motivation and engagement are seen as essential for cognitive processing to be effective (O’Donnell & Hmelo-Silver, 2013), just as the CoI framework emphasises the vital role of social presence and teaching presence to support cognitive presence (Garrison et al., 1999).

Studies in CSCL often use carefully designed, highly structured tasks (Vogel et al., 2021; Wise et al., 2021; Stahl et al., 2022) to allow comparison of different approaches to collaboration, commonly at the scale of a single classroom (Chen et al.,...
One approach that has been particularly fruitful in CSCL research is the idea of *scripting* the interactions between learners by assigning them to different roles within the task (Fischer et al., 2013; Vogel et al., 2021). Among groups of human participants, roles can be defined narrowly or more broadly and can be enforced more or less strictly (Rosé et al., 2008). Novice students might be given specific low-level tasks to carry out during a collaborative activity, such as entering the agreed answer into the interface (Olsen et al., 2017), while for more mature learners it is important not to “overscript” the expected collaborative interactions (Fischer et al., 2013; Vogel et al., 2021) but instead to leave room for learners to internalise the desired scripts so that they begin to use them autonomously.

Within CSCL, there are multiple views on what constitutes learning (Cress et al., 2021). The *subjective* view (Stahl & Hakkarainen, 2021), which aligns closely with the *cognitive* view (Chen & Teasley, 2022) and the epistemological stance of *individualism* (Ludvigsen et al., 2021), treats the learner as central, and sees learning as a change that takes place within individual minds, with collaboration providing input and stimulus (Oshima & Hoppe, 2021). Studies taking a cognitivist view tend to use quantitative, rather than qualitative, research designs (Borge & Rosé, 2021). In contrast, the *inter-subjective* view is concerned with how groups of people work together to solve problems collectively (Stahl & Hakkarainen, 2021; Chen & Teasley, 2022; Stahl et al., 2022), examining things like joint visual attention (Schneider & Pea, 2015) and physical engagement (Chen & Teasley, 2022), and holding that “group cognition cannot be reduced to individual learning.” (Chen et al., 2021, p. 169). The inter-subjective view is aligned with the epistemological stance of *relationism* (Ludvigsen et al., 2021) in its focus on the mechanisms of collaboration. Finally, the *inter-objective* approach considers external objects as places of learning, both physical and digital, where people interact with tools and with stimuli such as wiki-based platforms (Stahl & Hakkarainen, 2021; Chen & Teasley, 2022). Some studies that focus on technological advances in CSCL tools do not state an explicit epistemological position; they have been described as taking a “pragmatic and computational stance” (Ludvigsen et al., 2021, p. 53).

The notion of *transactivity* is an important one within the cognitivist tradition of CSCL (Suthers et al., 2013; Rosé et al., 2017). *Transactive contributions* are those that both demonstrate reasoning and also build on the earlier reasoning of other
contributors (Vogel et al., 2021; Chen & Teasley, 2022) – a definition closely resembling that of the Interactive mode in ICAP (Chi & Wylie, 2014). Transactive contributions have been shown to be correlated with learning (Rosé et al., 2017; Vogel et al., 2021). Contributions that demonstrate reasoning without building on anything that has gone before are labelled as externalizations (Rosé et al., 2008); the definition closely resembles the Constructive mode in ICAP. Within CSCL work on argumentation (Kimmerle et al., 2021), making reference to a previous contribution is considered to be the most important element of interaction and is given the name uptake (Suthers et al., 2010). The definition of uptake is intentionally broad; it includes replying to a message in a threaded discussion and also simply to “attending to another’s contribution” (Suthers et al., 2010, p. 14). In a similar way, the Integration phase of cognitive presence in the CoI framework is characterised by connecting ideas (Garrison et al., 1999); it does not distinguish between ideas based on reasoned arguments and ideas presented without justification. The work in this thesis is thus most closely aligned with the cognitivist tradition of CSCL.

In summary, this thesis contributes to the conceptual understanding of the cognitive quality of student participation in asynchronous online discussions. Our work explores how various attributes of the dialogue – linguistic and structural – relate to cognitive quality, represented by the CoI phases of cognitive presence and the ICAP modes of cognitive engagement. We found that some dialogue attributes were highly predictive for both CoI and ICAP, while others were only relevant to one framework. We also examined the relationships between the two theoretical frameworks directly, using machine learning and natural language processing to model cognitive quality. Looking at where the frameworks agree and where they differ revealed new insights into the factors that influence the cognitive quality of student participation, and provided empirical evidence that the two frameworks measure different aspects of interaction – and thus, that using both of them together provides a richer picture of student behaviour. We went on to use the two frameworks as a two-dimensional measure of cognitive quality, using a novel network analytic approach to quantify the associations between framework constructs. We considered the impact on those associations of two different instructional interventions, role assignment and external facilitation, described below (Section 1.2). Our final study looked at the possible effects of temporal ordering within the role assignment intervention.

1The work presented in this thesis includes verbatim copies of four peer-reviewed publications, produced under joint authorship with my supervisors. For consistency, the pronouns we, us, and our are used throughout.
1.1 Research questions

The overarching goal of this thesis was to propose a new methodological approach, combining insights from two theoretical frameworks to produce a comprehensive picture of the factors influencing the cognitive quality of student contributions to asynchronous online discussions. We wanted to discover what factors characterised high quality discussions, particularly where those factors might suggest concrete changes that students and instructors could make to improve the quality of group discussions and support the knowledge-building process. We also wanted to investigate and evaluate whether, and in what ways, the two instructional interventions present in our data set influenced cognitive quality.

Within the programme of work, the first research goal was to understand how the theorised properties of the CoI phases of cognitive presence and the ICAP modes of cognitive engagement were manifested in practice within the discussion dialogue. We posited that identifying common characteristics of messages that share the same label would provide valuable insights into the factors that relate to cognitive quality. Thus, our first research question was

**Research Question 1:**

*To what extent are theorised properties of cognitive quality associated with linguistic and structural attributes of dialogue in asynchronous online discussions?*

Since both the CoI and ICAP frameworks focus on cognitive aspects of learning, and both have been used as proxies for quality in prior work (Corich et al., 2004; Taskin et al., 2019), we wanted to compare them directly. The second goal of this thesis was thus to examine the relationships between the two sets of framework constructs, looking in particular at whether those relationships changed or remained stable in the context of different instructional interventions. Stable relationships between pairs of constructs could indicate that those constructs represented broadly the same thing and might be used interchangeably. Relationships that varied with the conditions could reveal nuances relating to the properties of those constructs. With this in mind, our second research question was
Chapter 1. Introduction

Research Question 2:
To what extent can a network analytic approach be used to examine the robustness of associations found between different theorised properties of cognitive quality under different instructional scaffolding conditions?

In addressing the second research question, we confirmed a large positive effect of the role assignment intervention on cognitive quality: discussion participants consistently reached higher levels of both the CoI phases of cognitive presence and ICAP modes of cognitive engagement while they were in the Research Expert role, compared to the role of Practising Researcher. When an intervention divides participants into groups that experience different treatments, it is important to ask whether one group thereby gains an advantage over another, particularly in the context of a graded assignment. One common way to limit the risk of harm is to rotate the participants through all the treatments in turn. However, one group may nevertheless gain an advantage as a result of the chosen ordering. Therefore, the third research question addressed by this thesis concerned the order in which participants experienced the two roles, and asked

Research Question 3:
To what extent can a network analytic approach account for the effects of temporal ordering within an instructional intervention experienced differently across student groups?

1.2 Methodology

The research presented in this thesis used a methodological approach grounded in data science, one of the three key dimensions of LA (Gašević et al., 2017). The set of asynchronous discussion forum messages used in this thesis was collected from an assignment that formed part of an existing course of study; data collection predated this thesis. Socio-cognitive research in CSCL sometimes involves collecting data from field studies (Hmelo-Silver & Jeong, 2021; Stahl & Hakkarainen, 2021); but it is also common for the collaborative tasks used in CSCL studies to be constructed specifically in order to measure and assess teamwork skills, without being linked to a course of study or genuine workplace activity, perhaps using computer applications designed specifically for collaboration (Stahl & Hakkarainen, 2021). In contrast, the open-ended assignment task from which our data set was collected (Gašević,
Adesope, et al., 2015) was not designed to answer our specific research questions and was carried out using a standard threaded discussion forum. Our work therefore follows a pattern that is common across many LA studies, where data traces from educational interactions among human participants, and between human participants and technological systems, are analysed after the fact in order to gain understanding of the effect of particular design choices or interventions (Siemens & Gašević, 2012; Gašević et al., 2017; Oshima & Hoppe, 2021).

**Supervised machine learning** We trained random forest classifiers to label previously unseen discussion forum messages with the framework constructs (the Col phases of cognitive presence and the ICAP modes of cognitive engagement). In previous work, random forests achieved better predictive performance than logistic regression and support vector machines (Yogev et al., 2018). Random forests also provide information about which of the model features had the greatest predictive power. Several such classifiers were trained to identify the relevant phase of cognitive presence and mode of cognitive engagement for each message in the discussion. In earlier work, the selection of appropriate features was found to be more important than the choice of classification algorithm when working with short texts (Rosé et al., 2008). The features used by the classifiers included linguistic and structural attributes of the dialogue. The linguistic attributes of the dialogue, such as word counts, text coherence measures, and complexity scores, were identified using two natural language processing tools, Linguistic Inquiry and Word Count (LIWC; Tausczik and Pennebaker, 2010) and Coh-Metrix (McNamara et al., 2014). We derived the structural attributes, including the depth of each message within its discussion thread and the number of replies it received, using Python scripts. The outputs of the classifiers were interpreted both quantitatively and qualitatively to determine how well they could assign labels to previously unseen messages, and to discover which of the model features were most informative.

**Network analytic approach** We adopted a network analytic approach that allowed us to consider both sets of framework constructs together, using Epistemic Network Analysis (ENA; Shaffer et al., 2016). ENA was developed in order to analyse data sets where the connections between labels are of more importance than the distributions of the labels themselves. In the present thesis, we are interested in the co-occurrences of labels from each framework. ENA enables comparisons between individuals and
groups within the same dynamic network model, providing both quantitative and qualitative insights. It reveals the pattern of connections between concepts and the relative strength of each connection. In this way, ENA could be seen as an extension to Correspondence Analysis (CA; Zhu and Bergner, 2017). Both CA and ENA start with a graph of connections and generate a visualisation in a lower-dimensional space; similar elements are clustered together to reveal patterns in the connections in a visual way. ENA plots can also include points indicating the weighted mean of the connections for a sub-network, such as that for an individual participant. The two-dimensional space is optimised to allow comparisons between such summary points in terms of the original data labels, supporting additional analyses. We combined ENA with quantitative data analysis in order to explore the associations between framework constructs, across and within frameworks. We also used ENA to compare sub-groups of participants.

ENA is a network analytic approach, described as “a basis for bridging over between data analysis and meaning-making” (Oshima & Hoppe, 2021, p. 579). ENA has been used in other LA and CSCL research (Knight et al., 2014; Rolim, Ferreira, et al., 2019; Hmelo-Silver & Jeong, 2021) and was useful for answering the research questions posed in this thesis. Alternative approaches and data analysis methods would yield different insights. For example, relational event modelling (REM; Schecter and Contractor, 2017) focuses on the behaviour of individuals, logging a relational event every time someone interacts with another person or an object; it can account for the tempo and pattern of exchanges between participants in ways that ENA cannot (Schecter & Contractor, 2017). An analysis using REM could investigate how the timing of exchanges affected cognitive quality: a quick response might suggest that the participant had not fully read and considered the previous message – scanning, rather than reading (Wise et al., 2014) – or could perhaps indicate that the reply was more likely to be a simple factual correction rather than a reasoned argument.

Another possible alternative approach is that of uptake analysis (Suthers et al., 2010), where contingency graphs are constructed to map possible paths of influence between actions and outputs of participants. Whereas ENA provides a population-level view based on pre-defined labels, uptake analysis can be used to trace the development of ideas within a discussion and across different threads. Uptake analysis can also handle situations that are beyond the scope of ENA, such as data from forums that allow message content to be co-authored by multiple participants or edited after posting (Rosé et al., 2017). Each successive version of the message content would appear as a separate entry in the contingency graph, along with dependencies between those
entries, between successive editing actions, and between the editing actions and the participant(s) who carried them out, enabling researchers to track the evolution of ideas and their representation over time.

It is also worthwhile to consider social network analysis (SNA). SNA looks at group dynamics in terms of the connections between data points in a network constructed from the messages that are exchanged between participants. Centrality measures are useful to reveal patterns of social ties and can be used to determine the role and influence of each individual (Joksimović et al., 2019). However, SNA does not take account of the labels assigned to the messages. In addition, the patterns of interactions in the data set we used were highly constrained by the parameters of the assigned task (described below), limiting the potential for variability in centrality measures. For this reason, SNA measures were not used in the work reported here.

The idea of units of analysis is central to ENA (Shaffer et al., 2016) and is widely used in CSCL (Suthers et al., 2013; Hmelo-Silver & Jeong, 2021; Stahl & Hakkarainen, 2021). Units of analysis can include physical actions and interactions, rhetorical moves, and written messages. For example, the unit of analysis in a study could be the set of messages sent by each participant in each thread. Aggregating the messages in this way allows certain measures to be calculated and compared per participant and per thread, such as average message length; but other information is lost, such as changes in message length over time within each thread – this is a well-known limitation of the so-called “coding and counting” approach (Rosé et al., 2017). In our analysis using ENA, we aggregated the messages sent by participants in several different ways. This allowed us to compare how the framework constructs varied across different instructional scaffolding conditions (described in the next section) and between groups differentiated by time. Thus, ENA was chosen for use in the studies reported in this thesis because of its central focus on the co-occurrences of labels and its ability to compare the behaviour of different sub-groups within a common frame of reference.

**Data set** The data used in the research described in this thesis came from a credit-bearing postgraduate course in software engineering and was collected in accordance with institutional ethics requirements. The course ran over six sessions and featured two separate instructional interventions. The data collection and choice of interventions predate this thesis. The work reported in this thesis was thus secondary data analysis (Suthers et al., 2013).
The first intervention, role assignment, was within subjects. Two distinct roles were defined, Research Expert and Practising Researcher, both of which were composite roles that incorporated several low-level duties, including contributing new ideas and making connections with what had already been said. Participants took on both roles in different discussion threads over the course of a discussion task that lasted several weeks, leading one thread as the Research Expert and contributing to other threads as a Practising Researcher.

The second intervention, external facilitation, was between subjects. It compared a Treatment group, who were given additional guidelines that aimed to enhance the quality of their self-regulation in online discussions (Gašević, Adesope, et al., 2015), and a Control group, who did not receive those additional guidelines. The additional guidelines gave examples of the type of high-quality questions participants were expected to ask while in the Practising Researcher role. The combination of the two instructional interventions thus defined four distinct instructional scaffolding conditions, with a similar proportion of messages in each condition:

- Practising Researcher in the Control group (24.8%)
- Research Expert in the Control group (23.5%)
- Practising Researcher in the Treatment group (26.7%)
- Research Expert in the Treatment group (25.0%)

The two interventions, role assignment and external facilitation, were selected by the original research team to support a “quasi-experimental mixed design” study (Gašević, Adesope, et al., 2015, p. 56). The roles of Practising Researcher and Research Expert, used in the role assignment intervention, were designed as composite roles with multiple functions in order to avoid the reported ineffectiveness of some single duty roles (Schellens et al., 2007; Gašević, Adesope, et al., 2015). Both roles offer participants the opportunity to become “members of a knowledge building community” (Scardamalia & Bereiter, 2005, p. 98). The literature regarding role assignment is discussed in more detail in Chapter 5.

The framing of the discussion task as a graded assignment and the guidelines given to all participants were informed by earlier work that looked at how to improve the quality of online discussions (Gilbert & Dabbagh, 2005; Rovai, 2007). Neither the CoI nor the ICAP framework was used in the development of the task guidelines; the participation requirements were not directly derived from either of the frameworks. The additional guidelines that were given only to the Treatment group were intended to reinforce the

---

2The guidelines are shown in Figure 1 in Farrow et al. (2021a), reproduced on p. 91.
1.2. Methodology

expectations of the instructors (Gašević, Adesope, et al., 2015) and thereby enable participants to “become more thoughtful judges of the quality of their work” (Rovai, 2007, p. 80).

Framework labels As mentioned above, the research in this thesis makes use of data which was collected before the thesis began and which formed the subject of previous research studies. During that earlier work using the same data set, each message was assigned a single label based on the CoI phases of cognitive presence (Gašević, Adesope, et al., 2015). For example, a message that proposed a relevant new idea would be labelled as Exploration, while one that connected several ideas together would be labelled as Integration.

In order to support the research goals of the present thesis, the messages were manually labelled once more, this time using an extended version of the ICAP framework. The manual coding scheme we developed, with rules for how to label the messages with the ICAP modes of cognitive engagement, was adapted from one used in prior work (Yoge, et al., 2018). For example, a message that paraphrased something that had already been said would be labelled as Active Targeted, while one that demonstrated reasoning in response to the content of a previous message would be labelled as Interactive. The messages were labelled by two independent annotators: the author of this thesis and another postgraduate research student, who was not otherwise involved in the research. The detailed coding guidelines we developed are included in Appendix A. The course design, the data set of messages, and the label definitions are described in more detail in Chapter 2.

Participant privacy In order to protect the privacy of the discussion participants during the manual data labelling process, we replaced all instances of personal names within the messages with pseudonyms. Using pseudonyms, instead of simply redacting the names, made it possible for annotators to identify messages where a participant was directly responding to a point made by another participant earlier in the discussion thread. We developed a semi-automated approach for replacing names with pseudonyms that took account of the message context. By assigning each participant a unique pseudonym, we ensured consistent handling of issues such as spelling errors and the use of nicknames. The pseudonymisation process we developed is described and evaluated in Chapter 2, where common privacy concerns surrounding the use of user-generated content are also addressed.
1.2.1 Specific methods used to address each research question

Our first research question (RQ1) asked about potential associations between the theorised properties of cognitive quality and the linguistic and structural attributes of asynchronous online discussions. In response to this question, we trained several predictive classifiers, using the dialogue attributes as model features (Chapter 3). We particularly wanted to discover what characterises high quality discussions, so the attributes that were found to be most discriminative as model features were investigated further, by looking at how they were correlated with the framework constructs. In this way, we discovered more about the characteristics of the different CoI phases of cognitive presence and ICAP modes of cognitive engagement, finding some commonalities between the two frameworks. We went on to train a second set of classifiers, using the construct labels as additional model features, providing us with another way to explore how the frameworks were related to one another and to the dialogue attributes.

To address our second research question (RQ2), we examined the patterns of association between the various constructs of cognitive quality. We used traditional cross-tabulations and heat maps to gain some initial insights. We developed a novel network analytic approach using ENA that allowed us to examine the framework constructs from CoI and ICAP together in the same analysis (Chapter 4). We looked at the data set as a whole and also considered the subsets of data defined by the two instructional interventions. We analysed label co-occurrences, using both traditional visualisations and ENA, to discover how the associations between pairs of constructs varied between instructional scaffolding conditions.

Our third research question (RQ3) concerned the role assignment intervention, where participants took on one of two complementary roles in each thread of the discussion. We used ENA to compare the effect of experiencing the Research Expert role at different points within the duration of the assignment. We constructed an ENA network from the whole data set and plotted the positions defined by the connections present in messages sent by participants in each group (Chapter 5). We evaluated three different definitions of role order and observed how the groups were positioned relative to the framework labels. The analysis used to address this question made use of the idea of comparing single networks as “snapshots” over time (Oshima & Hoppe, 2021, p. 572).
1.3 Thesis structure and overview

The studies that address the three research questions are organised into individual chapters of the thesis (Table 1.1). Each chapter is focused on one study, addressing one or more research questions. Chapter 2, Chapter 3, Chapter 4, and Chapter 5 each include a verbatim copy of a peer-reviewed publication. These chapters begin with a short overview of the publication, describing how it fits into the thesis structure, and conclude with a summary that indicates how the study results contribute to the overall research goals. Finally, in Chapter 6, we summarise the main findings and implications of the research reported in this thesis, and suggest some directions for future work. In the remainder of this section, we provide a brief overview of each of the chapters in turn, highlighting the research contributions and published research outputs relating to each of them.

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Research questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2</td>
<td>Data preparation</td>
<td></td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Markers of cognitive quality</td>
<td>✓</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Integrating multiple quality measures</td>
<td>✓</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Effects of temporal ordering</td>
<td>✓</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>Conclusions and future directions</td>
<td>✓ ✓ ✓</td>
</tr>
</tbody>
</table>

Table 1.1: Overview of the research questions addressed in each chapter.

1.3.1 Overview of Chapter 2: “Data preparation”

The use of asynchronous online discussion forums to support learning is now firmly established practice across diverse educational settings. Many theoretical frameworks have been developed to aid research into qualitative aspects of interactions in such discussions, with a view to encouraging deeper engagement with the course materials and concepts and thus improving learning outcomes. The empirical studies reported in this thesis examined the constructs of the CoI and ICAP frameworks in the context of a collection of messages that were exchanged between participants in a fully-online distance-learning course. Chapter 2 introduces the data set used in those studies and reports how we addressed ethical concerns by removing personal names from
the messages. It also describes how the data set was prepared for analysis by adding labels to the messages to indicate the relevant CoI phases of cognitive presence and ICAP modes of cognitive engagement.

**Research contributions:**

- We developed a novel approach for identifying personal names of participants within messages and replacing them with consistent pseudonyms, taking special care to handle nicknames and misspellings.
- We evaluated our pseudonymisation approach against a general-purpose tool for redacting personal information and showed that the tailored approach we developed, which takes into account message context, produced better results.
- We extended the ICAP taxonomy and adapted the labelling rules to reflect the setup of the discussion task in our data set. We developed a coding manual, which was used by two annotators to label the messages in our data set, obtaining substantial agreement.

**Research output:**


**1.3.2 Overview of Chapter 3: “Markers of cognitive quality” (RQ1)**

In previous work, random forest classifiers were used to label messages with the CoI phases of cognitive presence and to identify which attributes of the dialogue could be used to distinguish between them (Kovanović et al., 2016; Neto et al., 2018). No prior work had looked at linguistic and structural dialogue attributes in relation to the ICAP modes of cognitive engagement. In the studies reported in Chapter 3, we trained classifiers for both frameworks using the same set of dialogue attributes. We compared the features that were most discriminative for both frameworks to learn how cognitive quality is associated with attributes of the dialogue. Considering both frameworks together provided further insights into the factors underlying cognitive quality.
1.3. Thesis structure and overview

Research contributions:

- We identified linguistic and structural markers of cognitive quality in student contributions to asynchronous online discussions.
- We demonstrated that the CoI and ICAP frameworks measure different aspects of student behaviour, since several attributes of the dialogue were found to be predictive for only one of the two underlying frameworks.
- We discovered that both the CoI and ICAP frameworks attributed greater cognitive quality to messages that were more deeply nested in the threaded reply structure, suggesting that participants tended to write higher quality messages when adding to established message threads.

Research output:


1.3.3 Overview of Chapter 4: “Integrating multiple quality measures” (RQ2)

The concept of cognitive quality encompasses both cognitive presence, one of the three presences in CoI, and cognitive engagement, defined by ICAP. Chapter 4 looks at the associations between framework constructs and explores how the frameworks can be integrated to form a two-dimensional measure of cognitive quality. We also considered the possible moderating effects of the two instructional interventions that were used during the course. We developed a novel network analytic approach that enabled both of these matters to be addressed. We combined traditional cross-
tabulations and ENA in order to generate both quantitative and qualitative insights into the associations between framework constructs, and to assess how those associations were moderated by the instructional interventions. While some previous studies had looked at associations between multiple constructs – for example, between cognitive presence and the other CoI presences – it is more common in the literature for each construct to be addressed alone. To our knowledge, ours was the first study to use CoI and ICAP together in this way.

Research contributions:

• We presented a novel network analytic approach using ENA that allowed the constructs from multiple frameworks to be analysed together.
• As previously seen with other frameworks, we found that the labels relating to higher cognitive quality in both frameworks tended to co-occur, while the lower-level constructs were not so closely aligned.
• We found that the external facilitation intervention had a large moderating effect on the associations between the framework constructs, with the Treatment group scoring more highly than the Control group in terms of the CoI phases of cognitive presence but not the ICAP modes of cognitive engagement.
• In contrast, there was no clear moderating effect for the role assignment intervention. Measures of cognitive quality were consistently higher for participants in the Research Expert role than the Practising Researcher role.
• We found that the largest class within the CoI framework (Exploration) could be subdivided in a theoretically meaningful way by considering the ICAP label for the messages.

Research output:

1.3.4 Overview of Chapter 5: “Effects of temporal ordering” (RQ3)

The discussion task from which the data used in this thesis was collected made use of a role assignment intervention, where one of the roles (Research Expert) was limited to a single participant in each thread – the person who started the thread – and each participant only experienced that role once. Since the discussion task ran for several weeks, a different group of participants took on the Research Expert role each week. The study reported in Chapter 5 examined whether the chosen ordering might favour one group over another. Any time participants experience a within-subjects intervention over an extended period of time, it is unavoidable that some will experience the conditions in a different order from others. It becomes even more important to consider the possibility of disadvantage when the assignment contributes to the overall course grade.

This chapter builds on the work presented in Chapter 4. A similar network analytic approach was used to compare cognitive quality across groups. We considered whether the cognitive quality of the contributions changed over time; whether there was an effect on cognitive quality based on when each participant started their designated thread as Research Expert; and whether participants behaved differently before, during, and after the point when they started their own thread.

Research contributions:

- We used the network analytic approach that was developed in the previous chapter to answer questions about potential temporal ordering effects in the role assignment intervention.
- We found some small effects related to temporal ordering, with participants tending to contribute messages of higher cognitive quality to other threads in the week(s) after they had started their own designated threads.
- The results of this chapter confirm that the role assignment intervention had a large positive effect on cognitive quality overall, far outweighing the small effects due to temporal ordering, and did not confer an unfair advantage on any of the groups.
Research output:


1.3.5 Overview of Chapter 6: “Conclusions and future directions” (RQ1, RQ2, RQ3)

In the final chapter, we summarise the main contributions of this thesis and its impact in the context of the research questions identified in Section 1.1. We discuss the wider implications of the present work for research and practice, and conclude Chapter 6 with some suggestions for possible future research in this area.
Chapter 2

Data preparation

2.1 Introduction

Our research into the theorised properties of cognitive quality in online discussions was based on evidence gathered from a data set of messages exchanged in an asynchronous online discussion forum. This chapter introduces the data that was used in all the studies in this thesis. It describes the steps that were taken to prepare the data for analysis, in order to answer the research questions posed in Section 1.1. The data was collected with the consent of the discussion forum participants and in accordance with institutional ethics requirements. The data collection predates this thesis.

Since we do not have permission to publish the data itself or to share unredacted examples, the common features of the data that are relevant to all the studies reported in this thesis are described in the present chapter, beginning in Section 2.1.1 by describing the structure of the online discussions from which the messages were collected. There is some unavoidable repetition between the information presented in this chapter and in each of the published papers.

We added labels to each message in order to support the analysis conducted in the studies reported in Chapter 3, Chapter 4, and Chapter 5. Labels identifying linguistic attributes of the messages were added automatically using third-party tools: the LIWC software package, and the Coh-Metrix tool. Structural attributes, such as the number of replies each message received, were labelled using Python scripts. Labels indicating the CoI phases of cognitive presence and the ICAP modes of cognitive engagement were added by human annotators, based on the published descriptors of each of the framework constructs. The manual labelling process for both frameworks is described in Section 2.1.2.
Section 2.1.3 describes how we set out to build the foundation for our programme of research in an ethical manner, by removing the personal names of participants from the message text before they were shared with human annotators. These ethical considerations motivated our first published research study, presented in Section 2.2. Section 2.3 summarises the contributions of this chapter.

2.1.1 Structure of online discussions used in this thesis

The data used to investigate the research questions in this thesis came from a fully-online credit-bearing course in software engineering at Masters level that ran at a Canadian university over 6 sessions, from 2008–2011 (Gašević, Adesope, et al., 2015). The discussions took place in weeks 3–6 of each course session. Every student was expected to prepare and upload a video presentation based on a research paper of their own choosing, relevant to the course topics. Students shared a link to their presentation in a new thread and then led a discussion around the topic with their peers – instructors commented only rarely. The video presentations themselves were not made available for research, only the text-based discussions that followed. Student contributions to the discussion were graded and contributed 10% to the final course grade. A further 5% of the final grade was based on participation in unstructured course discussions on general topics, which were not included in the research data set.

Messages were written in English and were extracted for analysis as unformatted text. Before the message text was shared with annotators, all the personal names of participants were redacted from the text. The approach we used for the pseudonymisation step is described and evaluated in Section 2.2. All message examples given in this thesis use fictitious names.

A single message often contained several sentences, as participants reasoned about the topic and provided evidence for their statements. However, punctuation was not used consistently, reflecting the relatively informal nature of online discussion forums compared to written assignments. As a consequence of the more informal writing style, standard automated methods to divide messages into sentences were prone to failure and produced over-long run-on sentences, along with shorter incoherent fragments of text. In contrast, each individual message taken alone can be assumed to represent a coherent unit of contribution to the discussion (Garrison et al., 1999). For these reasons, analysis in this thesis was conducted at the level of individual messages,
as is common in the field (Garrison et al., 2001; Schellens et al., 2007; Kovanović et al., 2016; Hu, Donald, & Giacaman, 2021).

The user interface allowed messages to be added as replies to previously posted messages; thus, conversations were arranged in nested threads. A stylised illustration of the interactions between participants as they replied to one another’s messages during two sample message threads is shown in Figure 2.1.

In the example interactions shown in Figure 2.1(a), several participants responded to the opening message but did not interact with one another directly. The person who started the thread replied to each of them in turn. None of the exchanges continued beyond the initial question and answer. Message B was sent very soon after message A, so it is likely that those two participants did not see each other’s messages until after both were posted. The participant who sent message C had the opportunity to read all the earlier messages before sending their own message but may not in fact have done so. In contrast, it is reasonable to expect that the participant who started the thread and sent messages 1, 2, and 3 took into account the content of each message to which they were responding.

Figure 2.1(b) illustrates the interactions that took place in another discussion thread, which contained longer sub-threads and several examples of participants responding directly to one another and not only to the person who started the thread. Messages D and E were sent by two different participants in response to the opening message and quickly received replies (messages 1 and 2).\(^1\) A fourth participant continued the discussion from message 2 by sending message F. The final contributor to the sub-thread joined the conversation by replying to message F with message G\(_1\), also referencing message D in the text. The thread-starter replied to message F with message 3, and to message G\(_1\) with message 4.\(^2\) The sub-thread concluded when the same participant who sent message G\(_1\) replied to message 3 with message G\(_2\).

The threads in Figure 2.1 are illustrated separately, but it was generally the case that multiple threads overlapped in time, since the entire discussion task took place over the course of 3–4 weeks. Nevertheless, the content of each thread was typically self-contained; it was rare for participants to refer to topics from other threads, thereby simplifying the analysis by reducing ambiguity (Oshima & Hoppe, 2021).

\(^1\)Message E is shown partly occluded by message 2, which was sent soon afterwards.

\(^2\)Message 4 is mostly hidden by message G\(_2\) because they were sent so close together in time.
Figure 2.1: Patterns of interaction between participants in the threaded discussion forum, illustrated by showing the messages and replies sent by various participants within a single discussion thread. Solid arrows indicate a direct reply to a parent message. Dashed arrows indicate a reference in the message text to a message other than the immediate parent. The nodes representing messages sent by the person who started the thread are all coloured pink and numbered for reference. The rest of the nodes are arranged vertically, with one row per participant, and labelled with letters to identify the message author along with subscript numbers as needed to identify an individual message.
2.1. Introduction

2.1.2 Data annotation

The same set of messages that was used in this thesis as secondary data was previously used in several earlier studies of cognitive presence (Kovanović et al., 2014; Gašević, Adesope, et al., 2015; Waters et al., 2015; Kovanović et al., 2016; Farrow et al., 2019). In support of that earlier work, each message was labelled with its phase of cognitive presence by two expert annotators. The labels assigned to the messages enabled patterns in the data to be identified in general ways through analysis of label frequency and at a more detailed level by using computational modelling (Borge & Rosé, 2021). The annotators used the indicators of each of the four phases of cognitive presence (Table 2.1), and the socio-cognitive processes underlying them, defined by Garrison (2011). The Other label was used for messages that had no indicators of cognitive presence. After discussing how to apply the annotation scheme to one sample thread, the two annotators each labelled the messages from a second thread independently, achieving high levels of agreement (Gašević, Adesope, et al., 2015). The remaining messages were labelled independently, with differences reconciled through discussion once the labelling was complete. Inter-annotator agreement was measured as Cohen’s $\kappa = 0.974$.

The values of relevant linguistic dialogue attributes were also calculated: 91 word counts derived using the LIWC software package, and 106 metrics from Coh-Metrix, including text coherence, complexity, readability, and lexical category use. The studies reported in Chapter 3, Chapter 4, and Chapter 5 of this thesis made use of the labels indicating the CoI phases of cognitive presence. The studies reported in Chapter 3 also used the linguistic attributes that were assigned using LIWC and Coh-Metrix.\(^3\)

In order to address the specific research questions in this thesis, comparing aspects of the ICAP framework with the CoI framework, each message in the data set was additionally labelled with the ICAP modes of cognitive engagement. In this way, the secondary data was augmented specifically for the purposes of carrying out the research described in this thesis. Unlike the dialogue attributes that were calculated automatically by LIWC and Coh-Metrix, the framework constructs are defined in terms of indicator phrases and descriptors, and the labelling task required additional manual effort.

\(^3\)Additional features that were generated in the course of earlier studies, relating to the discussion structure, were not used directly but were recalculated as needed.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Descriptor</th>
<th>Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triggering event</td>
<td>Evocative</td>
<td>Recognize problem</td>
</tr>
<tr>
<td></td>
<td>(inductive)</td>
<td>Puzzlement</td>
</tr>
<tr>
<td>Exploration</td>
<td>Inquisitive</td>
<td>Divergence</td>
</tr>
<tr>
<td></td>
<td>(divergent)</td>
<td>Information exchange</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Suggestions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brainstorming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intuitive leaps</td>
</tr>
<tr>
<td>Integration</td>
<td>Tentative</td>
<td>Convergence</td>
</tr>
<tr>
<td></td>
<td>(convergent)</td>
<td>Synthesis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Solutions</td>
</tr>
<tr>
<td>Resolution</td>
<td>Committed</td>
<td>Apply</td>
</tr>
<tr>
<td></td>
<td>(deductive)</td>
<td>Test</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Defend</td>
</tr>
</tbody>
</table>
2.1. Introduction

<table>
<thead>
<tr>
<th>Label</th>
<th>Mode</th>
<th>Example behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Passive</td>
<td>Reading messages without responding</td>
</tr>
<tr>
<td>O</td>
<td>Off-task</td>
<td>Commenting without any relation to the current topic or the course</td>
</tr>
<tr>
<td>A2</td>
<td>Active General</td>
<td>Showing other signs of being engaged with course content</td>
</tr>
<tr>
<td>A1</td>
<td>Active Targeted</td>
<td>Referencing specific previous content</td>
</tr>
<tr>
<td>F</td>
<td>Affirmation</td>
<td>Affirming what was said in an earlier message</td>
</tr>
<tr>
<td>C2</td>
<td>Constructive Extending</td>
<td>Introducing new content to the discussion</td>
</tr>
<tr>
<td>C1</td>
<td>Constructive Reasoning</td>
<td>Displaying explanation or reasoning about the current topic</td>
</tr>
<tr>
<td>I</td>
<td>Interactive</td>
<td>Displaying explanation or reasoning about the current topic in response to an earlier Constructive message</td>
</tr>
</tbody>
</table>

Table 2.2: The extended cognitive engagement taxonomy used in this thesis, which we adapted from Yogevo et al. (2018), based on the ICAP framework.

Before sharing the messages with annotators who were not part of our research group, all personal names of discussion participants were replaced with pseudonyms using the approach described below, in Section 2.2. The pseudonymised messages were each labelled with one of the ICAP modes of cognitive engagement by two annotators working independently, using a process similar to that used in earlier work (Yogevo et al., 2018). We adapted the label definitions and the flow chart from that earlier work (Figure 2.2) to take account of the discussion context, specifically the video presentations. The extended taxonomy of modes of cognitive engagement is given in Table 2.2 and our revised flow chart is shown in Figure 2.3. Inter-annotator agreement was measured as Cohen’s $\kappa = 0.623$.

We added the Affirmation label to the taxonomy to capture messages of agreement and affirmation. In the earlier work (Yogevo et al., 2018), the label given to such messages depended on the label assigned to the previous message (steps 7 and 8 in Figure 2.2). Since we intended to automate the labelling process by training a classifier, it was important that the label for each message was defined independently. Otherwise, if an earlier message label needed to be revised, it could also require changes to the labels of later messages. By adding the Affirmation label to the taxonomy, all the
Figure 2.2: The extended ICAP annotation flow chart, reprinted from Yoge et al. (2018).
Used with permission of Association for Computing Machinery, from Proceedings of the Fifth Annual ACM Conference on Learning at Scale, Eran Yoge, Kobi Gal, David Karger, Marc T. Facciotti, and Michele Igo, Classifying and visualizing students' cognitive engagement in course readings, 1–10, 2018; permission conveyed through Copyright Clearance Center, Inc.
2.1. Introduction

Course content includes:
- course logistics
- lectures, presentations, reading lists, and assignments
- personal experience if directly related to course content
- but not personal introductions, motivation for taking the course (these are considered off-topic)

The current topic includes:
- the content discussed in the linked presentation of a paper
- the concepts, frameworks, experiments, results, and theories described in the paper
- the content of study mentioned in the introduction of the paper
- but not technical and stylistic aspects of the presentation

Explanation and reasoning includes:
- proposing an explanation, or a cause and effect relationship
- comparing or distinguishing between less or more conditions
- elaborating on a point made in the linked student presentation or in a previous post
- making a statement about the current topic and justifying it with evidence or personal examples
- making a statement or asking a question about the current topic, giving reasons why the commenter thinks this way

New content includes:
- a link to a student presentation
- information from the presented paper not mentioned previously
- a reference to other documents or resources about the current topic not (previously) mentioned
- a question, view, or comment that adds to the current topic, going beyond what has already been said in earlier posts, but not involving any explanation or reasoning.

Referencing content includes:
- seeking or answering clarification questions about specific points in the linked presentation or in earlier posts, without adding any new content
- paraphrasing or repeating something from the linked presentation or from earlier posts
- making connections between resources already mentioned

Signs of engagement include:
- seeking general questions about the current topic (e.g., “Is it useful?”)
- making non-specific references to earlier posts (e.g., “As others said”, “You’re welcome”)
- reporting a technical issue (e.g., “The font was too fuzzy.”)
- reporting a personal experience or event (e.g., “I read something that said”)
- continuing a previous on-topic conversation (e.g., “You’re welcome”)
- reporting a technical issue (e.g., “The font was too fuzzy.”)
- seeking administrative matters

Start

Does the post only talk about content that is totally unrelated to both the current topic and the course?

Yes

Does the post explain or reason about the current topic?

Yes

Does the post introduce new content to the discussion?

Yes

Does the post affirm what another user said in an earlier post?

Yes

Does the post reference specific previous content (spelling, paraphrasing, repeating, quoting)?

Yes

Does the post show any sign that the user is engaged with the course content (explicitly or implicitly)?

Yes

Extended ICAP annotation flow chart v1.8
Elaine Farrow 15 January 2020

Figure 2.3: The extended cognitive engagement annotation flow chart developed and used in this thesis. The short labels correspond to the extended set of ICAP modes of cognitive engagement, listed in Table 2.2.

messages could be labelled independently. If desired, a post-processing step could be run after the initial labelling pass, using the logic defined by Yoge et al. (2018) to change the label of each Affirmation message to a new one based on the label of its parent message.

The updated flow chart used in the present work is shown in Figure 2.3 and the full coding manual that was supplied to the annotators is provided as Appendix A. Human annotators needed to take account of the connections between messages and the content of earlier messages so that they could determine whether a message was introducing new content rather than paraphrasing or repeating something that had already been said. An important point to note is that a message can only be assigned the Interactive label if it is a response to the substantive content of an earlier message in the thread. The opening message in a discussion can never be Interactive. In the data set of messages used in this thesis, the opening message in each thread typically provided the link to the student’s presentation without any further content or discussion. Therefore, none of the replies to such opening messages could be considered Interactive unless they also referenced the content of a different message.

2.1.3 Ethical considerations

The importance of considering ethical issues in LA research has been recognised from the early days of the field (Meyer, 2004; Kitto & Knight, 2019). Researchers must always take care to avoid potential harm to study participants. One area of particular concern is personally identifying data. The need to respect and protect the privacy of individuals is recognised across many fields of research, and legal safeguards have been developed (Rudniy, 2018). Educational systems and LA studies should be designed in such a way that personal information is not inadvertently revealed to others (Flanagan & Ogata, 2017; Gursoy et al., 2017).

Some of the steps that can be taken to protect participants’ privacy are relatively straightforward: for example, it is good practice to store personal data separately from system data (Flanagan & Ogata, 2017), and to remove usernames from log files before analysis (Khalil & Ebner, 2016). However, it is much harder to disguise the identities of participants in an asynchronous online discussion (Yacobson et al., 2021). Participants may introduce themselves by name and sign their messages (Sun et al., 2019) as they build social connections within their group (Rolim, Ferreira Leite de Mello, et al., 2019). Participants also refer to one another by name in the messages they write (Rudniy,
For example, a discussion contribution might reference points made in earlier messages by using the names of the authors. It is important for human annotators to be able to identify the relevant earlier messages in order to distinguish repeated or paraphrased content from novel contributions to the discussion (Atapattu et al., 2019). For this reason, simply redacting all the names from the messages – for example, masking all names with a single replacement token like NAME – would be unhelpful.

In the publication included in the next section, we describe how we wanted to protect the identities of the discussion participants by replacing their personal names with pseudonyms in a consistent way, and to automate that process as much as possible. We also wanted to connect together different names for the same individual, such as nicknames – otherwise, a given name and a nickname would be replaced by unrelated pseudonyms, making it harder for annotators to track the flow of ideas between participants. Several participants in our data set shared the same first name and needed distinct pseudonyms. Since we did not find any existing tools that were able to handle these requirements, we developed our own, building on principles established in earlier work (Bosch et al., 2020). We evaluated our approach against a general-purpose tool for redacting personal information (Kleinberg et al., 2022). Practical tools are not always well-represented in LA and CSCL publications (Rosé & Dimitriadis, 2021), but we were keen to share both our thinking and the tool itself with the community.

2.2 Peer-reviewed publication: Names, Nicknames, and Spelling Errors: Protecting Participant Identity in Learning Analytics of Online Discussions

This section includes the verbatim copy of the following peer-reviewed publication, reprinted with permission:


The code is available on github at https://github.com/efarrow/nicknames.
Contributions: The ideas and analysis in the paper were developed and discussed between all authors of the work. The original idea, the experiments, and the writing were the work of the first author.
Names, Nicknames, and Spelling Errors: Protecting Participant Identity in Learning Analytics of Online Discussions


1 INTRODUCTION

It is common, and even educationally desirable, for contributors in online discussions to refer to one another by name and to sign their own posts [21, 22, 24]. Before using such data for research purposes in learning analytics, it is good ethical practice – and often a strict requirement [4, 10, 13, 18] – that personally identifying information (PII) is removed. The category of PII is not limited to names and also includes email addresses, phone numbers, user names, dates of birth, places of work or study, and other pieces of data that could be used to identify an individual [11]. The content of PII is generally of little interest to educational researchers, who have no need for private information such as dates of birth. In fact, removing PII can be beneficial for analysis, since it adds unwanted noise to metrics like word and sentence length, particularly for very short messages. Personal names require careful handling. While metadata can be removed and other elements of PII can simply be redacted, personal names are often used to indicate the intended recipient of a message and to refer back to points raised by others in earlier messages. Masking, where a single replacement token (e.g., NAME) is used to redact all names throughout the data set [13] might be sufficient for some use cases [17, 23], but it discards important information [19] and can harm performance on subsequent analysis tasks [3]. In order to identify the same individual across different messages, personal names must instead be replaced consistently with alternative identifiers, or pseudonyms. Additionally, variant forms of the same name must be grouped together, and individuals with similar names must remain distinct.

In other work, the task of tracking mentions of the same individuals throughout a text is often carried out using coreference resolution [20]. Coreference resolution identifies the most likely connections between proper names and references such as pronouns (e.g., she) and expressions (e.g., the author)1. In contrast, the approach described in this work does not deal with pronouns or general referring expressions at all. Instead, we focus on full and shortened forms of proper names, along with misspellings. Our approach is not intended to replace coreference resolution but is, instead, a pre-processing step. In particular, pronouns and definite noun phrases (e.g., "the course instructor") are not replaced, since they cannot be used alone to identify specific people. If additional information is made available alongside the transcripts, such as the course title, dates, or time stamps, it might become possible to

\[1\] For example, in “Robert said that he had read the book,” the word he is likely to refer to Robert; whereas in “Robert asked if he had read the book,” it is more likely that he refers to someone else.
re-identify individuals from such descriptive phrases [25]; further anonymisation effort would then be required [13, 25].

The process of manually identifying and replacing personal names can be time-consuming and error-prone [25]. We developed a semi-automated approach to the task of identifying personal names and replacing them with alternative identifiers. We applied it successfully to a data set of messages collected from a distance-learning course. We evaluated the output with reference to the final processed data, in order to determine the importance of handling elements such as misspellings. We found that a relatively simple approach using regular expressions worked better than one that was more computationally demanding, without requiring any additional data to be annotated. Our approach can be adapted easily to handle a wide variety of data sets with differing characteristics.

The main contributions of this work are 1) to highlight the challenges involved in replacing personal names with pseudonyms in a consistent way across a corpus of informal written messages, when state-of-the-art methods may perform poorly on this type of data, and 2) to introduce a semi-automated approach that was developed specifically for online discussion forum messages and has been used successfully to pseudonymise a data set of such messages; and 3) to investigate the relative frequency – and thus importance – of different categories of personal name found in discussion forum messages, including shortened names and misspellings.

2 RELATED WORK

Ethical concerns within learning analytics have tended to focus on avoiding potential harms to learners and other research participants [14]. Learning management systems increasingly employ “anonymity by design” [3], storing learners’ personal data separately from system usage data, such as log files, that are commonly exported for use in learning analytics. Statistical disclosure controls preserve privacy by adding random noise to the results of statistical queries, such as counts and averages [10]. However, such measures do not account for elements of personal data that may be present in user-generated content [25]. Valid concerns around participants’ privacy mean that discussion forum messages are often excluded from the published data extracted from MOOCs [9]. Such messages may reveal private thoughts and opinions which participants choose to share selectively, with peers, but not with the wider public [7]. Yet, even when researchers have permission to access the raw data in full, it is still necessary to remove personal details before sharing the data with others, such as paid annotators. In the case of student lab reports, for example, simply removing the metadata that links a report document with the author’s username would not be sufficient to hide their identity, since learners often refer to themselves and other individuals by name in the body of the text [23].

Much of the work on identifying and removing personal names comes from the medical domain [1] and is focused on anonymisation through redaction. In a study using clinical data, a database of 3.8 million names collected from Social Security data was used to identify and redact personal names from narrative reports written by medical personnel [12]. The size of the database of valid names played an important role in system performance. Competing systems were seen to improve after they were given access to the larger list. However, since many rare names overlap with common dictionary words, between half and two-thirds of all tokens could end up removed from the reports unless the system also made use of frequency counts when deciding whether a word was a name.

Written personal exchanges between individuals, such as email messages, often contain sensitive details that need to be obscured before they can be used for research purposes. A corpus of approximately 2,500 personal email messages was pseudonymised using a hybrid approach combining semi-supervised and manual steps [19]. Words and phrases that were considered to represent sensitive data, including names of people and corporations, were replaced with alternative values. A notable feature of the study [19] was that the replacement names were specifically chosen to preserve the “nature” of the original names – for example, companies of a similar type. Pseudonyms were substituted for sensitive terms consistently across the corpus, indicating that duplicated personal names received no special handling.

Multimodal data poses additional challenges. In a corpus of Dutch Sign Language, names were removed from the textual annotations, but the original video was left unchanged, while in a corpus of German Sign Language, names were also removed from the video, by superimposing black rectangles over the relevant parts of the image [11]. In order to replace personal names in audio or video data with alternative names, it would be necessary to re-record the segment. However, text-based transcriptions and annotations can be treated in the same way as other textual corpora.

Named entity recognition (NER) software seems like an obvious choice for detecting names in discussion transcripts, and was used successfully to anonymise a corpus of chat logs in six languages [1]. However, issues like spelling and grammatical errors – common in informal texts – can dramatically reduce its effectiveness [19]. A brute-force approach, which simply looked for the names in the class register, out-performed two different NER implementations, in a study assessing how well the personal names of learners and instructors could be redacted from a data set of 1,000 student lab reports [23]. Another recent study, using a corpus of discussion forum text data from two online courses [4], found that a NER-based approach was not satisfactory for redacting names. Instead, the authors developed bespoke text anonymisation software that used machine learning to classify possible names. Their approach had three main stages:

(1) Identify possible name words.
(2) Classify the words as names or non-names, either manually or using machine learning.
(3) Remove from the text all identified names.

Recent advances in the area of Natural Language Processing (NLP) have led to the widespread use of pre-trained language models such as BERT [5] for many tasks. Model training makes use of enormous data sets and large amounts of computing power to train on low-level downstream tasks, using a much smaller amount of training data. In service of text anonymisation, the Textwash tool [15] used a BERT model, fine-tuned on annotated data from the British National Corpus [3], the Enron email corpus [16], and Wikipedia, in order to

https://github.com/efarrow/nicknames
identify PII entities. In addition to names, Textwash also redacted locations, occupations, dates and times, and many other classes of information that could be used to identify someone. The evaluation of Textwash was unusually robust, focusing on the likelihood of de-anonymisation of individuals in a realistic setting. Famous people could often be re-identified based on small pieces of information such as roles they had played in movies, while less famous people were almost never de-anonymised.

The present work aims to address several weaknesses of earlier approaches, while presenting a novel process that can be used to pseudonymise messages from asynchronous online discussion forums without requiring additional data annotation. There is growing interest within learning analytics in the analysis of online discussions [6, 7], but valid ethical concerns remain, relating to protecting the identities of participants, even where studies are conducted in-house. Additionally, restrictions on data sharing frequently hamper the reproduction of results. Many of these concerns could be alleviated by reliable pseudonymisation methods. Our first research question was thus:

RQ1: How well can regular expressions identify the personal names in discussion forum messages and connect them to the correct participants, compared to using the class list or a deep neural network?

We also explored the many-to-many relationship between people and names, an aspect that is often overlooked, leading tools to treat every instance of a given name as a reference to the same individual. Nicknames, misspellings, and similar artifacts are common in informal written texts but tend not to appear in the curated sources on which many NLP tools are trained. For this reason, standard tools may not handle informal texts well. The second research question addressed in the present study was therefore:

RQ2: What is the impact of non-standard names, such as nicknames and misspellings, on the task of pseudonymising discussion forum messages?

3 SCOPE OF STUDY: CATEGORIES OF PERSONAL NAMES

Personal names can take many forms. It will generally be impossible to predict all possible name variations that could be used in an informal online discussion, even when a full list of participants is available. Instead, it is necessary to take a data-driven approach. This section addresses some of the common issues we encountered that made personal name identification difficult. The example discussion in Figure 1 illustrates how the personal names used in messages should be replaced consistently with pseudonyms, despite spelling errors and other variations.

Full names of participants may be available, particularly where the data is collected during a course of education. This is a useful starting point for pseudonymisation, but will directly cover only a few cases in an informal discussion forum. Students will often be registered under their legal names but they may use a different name in everyday situations. Full names often have several parts (e.g., first, middle, last), of which a subset might be used in messages – perhaps just the given name, which could be a middle name. Some given names have more than one word (e.g., Mary Jane) but must be replaced with a single copy of the pseudonym (e.g., [U43]).

Shortened forms of names can exhibit a lot of variation. For example, an individual whose first name is Robert might sign messages as Rob, Robbert, or Bob (among others). Another common choice is to use initials (e.g., Jr or R.J.). Nicknames may not be related to any of the parts of the participant’s full name, but must nevertheless be redacted to preserve the privacy of the participant. Depending on the research goal, shortened names and nicknames might be allocated their own, related, pseudonyms.

Misspelled names are common in informal text-based discussions. Some misspellings are simple typos (Margret for Margaret), mistaken capitalisation (R0bert for Robert), or using only part of a multi-part name (Mary for Mary Jane). Common alternative spellings may be substituted (Elizabeth for Elisabeth). Stylised forms of names may be used to sign messages, such as alternating letters and spaces (K o b e r t), or surrounding initials with dashes ([R-G]). Misspellings should normally be replaced in the same way as the correct name (e.g., using [U12] for both Arthur and Arch). Duplicate names arise when two or more participants in a discussion forum share a name. Where the downstream research tasks require individuals to be traced across messages, it will be necessary to disambiguate such duplicate names so that the correct pseudonym can be assigned. Often this can be achieved automatically, simply by paying attention to scope and context, e.g., whether the two participants joined the course in different years. If the participants contribute to the same thread, one or both of them may (temporarily) alter the way they sign messages to avoid ambiguity, increasing the overall number of names in use.

Glued words can disguise names. Names should generally only be replaced when they appear as full words, to avoid phrases like “a summary report” becoming “a sum[U43] report”. However, care needs to be taken to check for words that have been accidentally glued together, for example, “thanksMary”. This issue is often encountered when sentences are run together, if the tokenisation relies on white space.4

Unwanted matches arise where a personal name does not refer to a conversation participant and should thus not be replaced, for example, names of public figures. Leaving the names of public figures unchanged is often desirable – for example, so that they can be discovered by a named entity recogniser in subsequent research tasks. It would be wrong to substitute the pseudonym [U12], relating to the participant named Archer, in a reference to the author Arthur C. Clarke. Such a substitution would increase the risk that the mapping from the pseudonym to the real name could be deduced and the participant’s identity revealed.

4 METHOD

Our approach to redacting personal names followed a similar three-step approach to earlier work by Bosch and colleagues [4].

3For charity, in the examples in this paper we use identifiers of the form [U43] as replacements – categorisation, rather than true pseudonymisation [31, 19]. A second substitution step could replace the identifiers with alternative names, if desired.

4One of the sample outputs from Textwash demonstrates how easily names can be missed when the input text is not well-formed according to the assumptions of the tokenisation tool: “a really good song MORE did with MORE evel is like MORE...” [15, p. 12].
(1) Automatically identify candidate name words.
(2) Filter out non-names and add missing names.
(3) Perform the substitution.

There are two main differences between the approach of Bosch and colleagues [4] and ours. First, the goal of the earlier work was basic anonymisation, so all confirmed names were simply redacted. In contrast, our aim was to track individuals across messages and replace their names with (unique) pseudonyms, to support later analysis of the conversations. Second, the prior work used a set-theoretic approach to identify all possible name words in their corpus, by removing any dictionary word that was not also a known name or location. However, a name identified in isolation would not be immediately useful for our task. Instead, we used regular expressions to collect candidate names for each participant (Section 4.2) and maintained a mapping between the names identified in the data and the participants who used those names. Our approach allowed for both manual and automated refinement of the mapping before the substitution step. Figure 2 shows a schematic diagram of the three-step process.

We evaluated our approach on a data set of messages collected from a distance-learning course. Similar to the lab reports in earlier work [25], the messages in the present study contained both personal names, to be pseudonymised, and names of cited authors, to be left unchanged. We conducted a post hoc evaluation of our approach with reference to the complete list of personal names found in the data set. The evaluation metrics included recall, precision, and $F_1$ scores; the number of individuals where all names used for them were replaced; and the total number of missed connections between an individual and a name. These metrics were computed for the candidate names identified by regular expressions and compared to a baseline using only names derived from the official class list. A second comparison used the set of names identified by the supervised machine learning model from the Textwash system. Textwash can identify several different classes of PII. The evaluation included only the tokens that Textwash labelled as PERSON_FIRSTNAME or PERSON_LASTNAME.

The rest of this section is organised as follows. The data set we used for evaluation contained examples of personal names in each of the categories described in Section 3. We briefly describe the data and the course from which it was collected in Section 4.1. The three stages of our approach are presented in detail: identifying possible names in the data (Section 4.2); filtering out non-names and adding missing names (Section 4.3); and performing the substitution (Section 4.4). The post hoc evaluation we carried out in order to answer our research questions is outlined in Section 4.5.

4.1 Data Used in the Study

The data we used was collected across 6 sessions of a Masters-level distance-learning course run by a Canadian university (Table 1). We needed to remove the personal names before sharing the discussion messages with annotators who were not part of our core research group. There were 84 unique participants in the data set and the post hoc evaluation revealed that they used 148 unique personal names between them, including multi-word names. Since some names were shared by multiple people, the total number of valid connections between a person and a name was 163.

The participation instructions for the discussion assignment indicated that every student should start a new thread, in which they would share a video presentation and field questions and comments from their peers. Course instructors rarely took part in the discussions. The user interface allowed messages to be added as replies to previously posted messages; thus, conversations were arranged in nested threads. Messages were written in English and were extracted for analysis as unformatted text. Consent was obtained to collect and analyse the data, in accordance with the requirements of our institution, but we do not have permission to share the data; example sentences in this work are illustrative.

The metadata associated with each message included the time stamp when the message was posted, along with numerical identifiers for the following:

- the session in which the course ran;
- the thread within the session;
- the participant who posted the message;
- the message itself; and
- the parent message, if any (zero for the top-level messages that started each thread).

We augmented the metadata by adding a derived field to indicate the identity of the participant who posted the parent message, if any, since that individual was the person most likely to be addressed by name in a reply. The value of the derived field was set to zero for top-level messages with no parent.
The steps we took to process the data and replace the personal names with alternative identifiers using our three-step approach (Figure 2) are presented next.

### 4.2 Step 1: Identifying Candidate Names

The first step towards replacing personal names with pseudonyms consistently across the data set was to identify a collection of strings that were likely to be personal names that needed to be replaced. After looking at a sample of our data, we chose to use regular expressions to collect candidate names from the messages. Note that we did not need to collect every instance of a name, but only the set of names used for each participant. We also did not need to collect every name, since many names referred to the authors of published research papers that were being discussed, rather than to the participants themselves.

To identify possible names used by the sender of the message, we looked at the end of each message for the regular expression `\w+(-\w+)?(\s+\w\.)?`. This can be expanded for ease of reading as ‘WORD CHAR’, ‘WORD CHAR’, or ‘WORD’, where CHAR indicates a Unicode word character and WORD indicates a sequence of one or more consecutive CHARs, with an optional hyphen in the middle. We hoped that this regular expression would capture sufficiently many examples of participants signing their messages, while at the same time avoiding excess noise. All sequences of text that matched the regular expression were treated as potential names. The trailing punctuation, if any, was stripped off, and any match that contained a decimal digit was discarded.

To find candidate names for message recipients, we looked at the start of each message for the case-insensitive phrase ‘Hi WORD’, using the regular expression `^[hH][iI]\W+\w+(-\w+)?`. The text that matched WORD was returned as the search result, and we again dropped any result containing a digit. We made the working assumption that every reply was addressed to the participant who posted the parent message, and connected candidate names with individuals on that basis.

A single set of candidate names was generated for each participant, combining the names identified in their roles as sender and (assumed) recipient of messages. We kept track of how many messages contained each name — a lower bound on the total number of instances of each name in the data, since it ignored multiple uses of the same name within one message. These counts allowed us to rank the names identified for each participant by their frequency of use.
We generated two additional name mappings to support the evaluation of our approach. For our baseline, we used the names from the class list and generated additional candidate names by splitting each full name into shorter forms composed of subsets of its parts (e.g., first and last, first and middle, each part alone). Names from the class list that did not appear anywhere in the data were removed. For a more challenging comparison, we collected the personal names that were identified by the Textwash model and connected them to both the sender and (assumed) recipient of the message in which they appeared.

4.3 Step 2: Filtering and Refining the Names

In order to allow the mapping from participants to names to be filtered and refined, it was written out to a text file in a simple, human-readable format (Figure 3). The mapping file was designed to be edited, manually or automatically, before being used to replace the names in the data. Obvious non-names can be removed and additional entries can be added. For example, for participant [U81], both Thanks Robert and html would be removed. In future, the removal of non-names could perhaps even be automated, using a machine learning approach like that in earlier work [4].

Mistaken connections between participants and names also need to be removed. Sometimes participants add their messages at the wrong level in the nested thread structure. For example, a reply addressed to one participant might appear nested under an earlier reply sent by another participant. In cases like that, the simple heuristic used to identify the likely recipient of a message would connect the name with the wrong participant (e.g., the name Maggie in the list for participant [U83]). If a class list is available, it can be used to identify names that are out of place. Shared cultural knowledge might also suggest that some names belong together (grouping Maggie with Margaret). Our system tracked how often each name was connected to an individual and displayed the names in frequency order. In the absence of a class list, frequency information could be used to determine which names were most likely to relate to each participant.

When a glued word appears in the list of candidate names (e.g., thanksMary), the glued part of the word (thanks) will be dropped during the substitution. Alternatively, a separate entry could be added, connecting the glued word with a composite output (e.g., thanks [U43]). A third option is to delete the glued word from the mapping and edit the corpus data file directly.

It may be desirable to replace multi-word names with a single token, for the benefit of later analysis. The name Mary Jane would otherwise become [U43] [U43], while the shortened form MJ would become [U44]. In addition, the likelihood of encountering duplicate names within the same context is lower for multi-part names like Mary Jane than for single word names like Robert. Since the regular expressions described in Section 4.2 return only single words as candidate names, or a word followed by a single letter, we augmented the mapping file by adding all the candidate names generated from the class list. Additional full names could also be added manually; for example, the name Robbie Jones might be added for participant [U81], based on the high frequency of occurrence of those names individually.

By comparison, the Textwash system automatically concatenates adjacent words found in the text, creating multi-word tokens from consecutive words that are identified as being the same type of entity. A multi-part first name like Mary Jane would thus be recognized as a single name. However, since Textwash uses different entity types for first names and last names, full names are never completely rejoined. This limitation means that a system relying on Textwash to identify possible names would replace a single mention of the name Elizabeth Brown with two consecutive instances of the token [U80]. In addition, names that are in common use as both first and last names, such as Arthur, may not be merged as expected.

Rather than adding multi-word names to the mapping file, an alternative approach would be to use only single-word names in the initial substitution and then to compress consecutive instances of the same token into a single instance in a post-processing step. While simpler in some ways, the resulting increase in duplicate names requiring manual resolution might make such a two-step approach undesirable.

4.4 Step 3: Substituting Personal Names

The filtered mapping file was used to replace the identified names with their associated pseudonyms consistently across the full data set. A regular expression was generated from each candidate name: `\b\w*\b\w*\b`, where NAME was the properly escaped form of the name. Wrapping the name in this way ensured that only full words were replaced. The names were sorted and substituted longest-first, so that multi-part names would be found before their constituent parts and could be replaced with a single instance of the correct pseudonym.

To deal with duplicate names shared by two or more individuals, we experimented with different ways of grouping the messages while carrying out the substitution step: one discussion thread at a time, one course session at a time, and the full data set at once. Where the same candidate name was connected to multiple participants within the same group (e.g., the name Robert for both [U81] and [U84]), a warning message was generated by the system so that the conflict could be resolved manually. It is worth noting that the same naming mapping file was used in every case—we there is no need to create specific mappings for subsets of the data.

4.5 Post Hoc Evaluation

After the name replacement step had run, the substitutions were reviewed by the research team. Names shared by multiple participants (e.g., Robert) and flagged by the system as ambiguous were manually assigned to the correct individual. During the review, the team discovered a small number of examples of personal names that had been missed. It was easy to add additional entries to the mapping file to connect the names with appropriate identifiers and then rerun the name substitution step to catch all instances of the same name.

Unwanted substitutions were reverted, such as Arthur C. Clarke in the example described in Section 3. While it would be possible to identify some public figures automatically from sources such as Wikipedia, the non-participant names we encountered were more often those of authors of scientific papers, most of whom are unlikely to be listed in Wikipedia. If desired, names that should be
left unchanged could be added to the mapping at the review stage, and the substitution algorithm updated to handle them accordingly.

For the evaluation, we referred to the final, corrected version of the data set and created a gold-standard mapping file, connecting all the personal names found in the data with the correct participant identifiers. This gold-standard mapping was used to evaluate the mapping generated using regular expressions and to compare it against a baseline that only used the class list, and against the mapping generated using the Textwash model, addressing RQ1.

Recall is arguably the most important metric for this task [2], since it indicates what proportion of the personal names in the data were correctly identified. Missing even one name means that the identity of the individual could be revealed [23]. Where the missed name in a message is a single character, the risk of re-identification is low, but the direct link is nonetheless broken between that message and others where the same individual is mentioned. Precision indicates what proportion of the suggested mappings are correct. Precision can be improved by removing non-names and mistaken connections (Section 4.3). If left uncorrected, low precision in the mapping will lead to a higher incidence of wrong substitutions that need to be reverted. The $F_1$ score, which is defined as the harmonic mean of recall and precision, is included for completeness.

In addition to these standard metrics, we defined two more, specific to the pseudonymisation task: missed connections and coverage. Since some names were used by multiple people (like Robert in the example), simply counting the found names could miss cases where a name was correctly connected to one individual but not to another. Instead, we counted the number of missed connections between a person and a name, compared to the total number of valid connections in the gold-standard mapping. Similarly, we calculated the coverage given by a mapping as the proportion of participants where every name used for that individual was correctly identified in the mapping.

The complexity of the name replacement task in general depends to a large extent on the number of unique names used for each individual and the difficulty in identifying those names. For example, variants like Elisabeth and Elizabeth.

We note that the same name can be a valid name for one participant and a misspelled name for another; for example, variants like Elisabeth and Elizabeth.

The post hoc evaluation concluded with further system comparisons. We looked at the manual edits and deletions required in the filtering step for both the regular expressions and for Textwash. We carried out an error analysis on the missed connections. Finally, we identified the optimal scope to use while grouping messages in the substitution step — session, thread, or whole data set — in order to capture names shared by multiple individuals.

### 5 RESULTS

#### 5.1 Initial Mapping from Participants to Candidate Names

We compared the gold-standard mapping from participants to candidate names against the mapping generated by using regular expressions, a baseline mapping using names from the class list, and the mapping generated using the Textwash model (Table 2). The table also includes results for the mappings generated by combining the class list with the other two approaches, for cases where a class list is available. Many of the incorrect connections in the initial mapping were removed during the filtering step (Section 4.3); they are included in the results in this section to allow a fair comparison of the candidate identification methods.

On inspection of the missed connections, there were several examples where the target was a multi-word name and the mapping correctly contained all the individual words; for example, a participant signing a message as Robbie Jones. We therefore carried out a further comparison after splitting the multi-word tokens in all the mappings into single words (Table 3). The splitting operation resulted in 159 unique single-word names and 169 connections between individuals and names.

In answer to RQ1, the regular expressions performed better on every metric compared with the deep neural network model from Textwash. The regular expressions achieved better coverage and higher recall and $F_1$, compared to using the class list alone, but at the expense of lower precision. The best coverage and recall scores were achieved by adding the names from the class list to the mappings extracted using regular expressions; this combination also gave the best $F_1$ score in Table 2, but there was a reduction in precision compared to the regular expressions alone. When only single-word names were considered (Table 3), both precision and $F_1$ score were lower.
To address RQ2, we quantified the prevalence of different types of personal names in the manually corrected data, where all personal names were consistently replaced by pseudonyms (2,888 substitutions in total). We found that the registered names of the participants accounted for more than half of the connections and most of the substitutions, although full names were rare (Table 4). Nicknames and other shortened or stylised forms were common and accounted for 23.2% of substitutions. Misspellings tended not to be repeated, although the same name could generate several different misspellings. The distribution of participants’ names across categories is shown in Figure 4.

### 5.2 Prevalence of Name Variants and Misspellings

To address RQ2, we quantified the prevalence of different types of personal names in the manually corrected data, where all personal names were consistently replaced by pseudonyms (2,888 substitutions in total). We found that the registered names of the participants accounted for more than half of the connections and most of the substitutions, although full names were rare (Table 4). Nicknames and other shortened or stylised forms were common and accounted for 23.2% of substitutions. Misspellings tended not to be repeated, although the same name could generate several different misspellings. The distribution of participants’ names across categories is shown in Figure 4.

### 5.3 Manual Edits to Remove Non-Names and Mistaken Connections

Some of the candidate names collected by the regular expressions were clearly not names at all entries like all from ‘Hi all’, Cheers from a sign-off, and Again from a message that began “Hi Again this is a question”. These are easily removed, as discussed in Section 4.3. Mistaken connections between individuals and names are harder to resolve. In practice, the largest set of possible names suggested for an individual participant in our data set only contained 20 entries, so the task of filtering them manually to remove mismatches was acceptably fast – particularly as we had a class list to guide us. Frequency data is also informative: a name that is connected to one participant many times but only once to another participant may be the result of non-standard message nesting (Section 4.3).

We noticed a small number of cases where a personal name had clearly been run together with an adjacent word to form a glued word (e.g., thanksMary). Our chosen remedy in this case was to edit the corpus data file directly to insert a space, and to remove the glued word from the list of candidate names. We also encountered several examples of people signing off with just an initial, picked up by the regular expressions as Thanks X or Cheers X. We removed those phrases from the list of candidate names and added the initial X instead. The most common unwanted substitution in our data was where a participant signed off using just the initial G. In addition to being used as a middle initial in an author’s name, the value G appeared in a formula in several messages.

Surprisingly, we saw similar problems with the names identified by the BERT-based model used in Textwash as with those collected by regular expressions. There were many examples of phrases like Hi Bob and Thanks Mary being wrongly identified as names, along with words like Thanks and Great. Some of the words identified as names were not personal names at all, but instead names of technologies. Additionally, since there was no way to inform Textwash to select only the names of the actual participants in the discussion, references to research papers also yielded the authors’ names.

The recall score for the names found by Textwash was lower than the recall for the regular expressions (Table 5) but higher than the recall of the class list alone. However, the very low precision of the Textwash results, below 6% even after the addition of the class list, indicates that it would not be a good solution in practice. The amount of manual effort required to remove the mistaken connections between candidate names and participants would be prohibitive.

### 5.4 Missed Connections and Missed Names

Using the mapping that was generated automatically from the regular expressions, 28 of 163 possible connections between names and discussion participants were missed (Table 5). The largest group of these (10 examples) were multi-word names, all of whose individual parts were present in the mapping – meaning that the substitution step would replace each name with multiple copies of the correct replacement token. Although this is not the desired output, the participants’ identities would not be revealed. There were 5 additional names where the only missing part was present in the class list.

### Table 2: Initial Mapping Generated by Each Approach, Compared Against the Gold-Standard Mapping

<table>
<thead>
<tr>
<th>Approach</th>
<th>Coverage</th>
<th>Missed Connections</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular Expressions</td>
<td>71.4%</td>
<td>28/163</td>
<td>82.8%</td>
<td>52.1%</td>
<td>64.0%</td>
</tr>
<tr>
<td>Class List</td>
<td>41.7%</td>
<td>80/163</td>
<td>50.9%</td>
<td>83.0%</td>
<td>63.1%</td>
</tr>
<tr>
<td>Textwash</td>
<td>59.5%</td>
<td>43/163</td>
<td>73.6%</td>
<td>3.8%</td>
<td>7.2%</td>
</tr>
<tr>
<td>Regular Expressions + Class List</td>
<td>79.8%</td>
<td>19/163</td>
<td>88.3%</td>
<td>51.1%</td>
<td>64.7%</td>
</tr>
<tr>
<td>Textwash + Class List</td>
<td>65.5%</td>
<td>35/163</td>
<td>78.5%</td>
<td>4.0%</td>
<td>7.7%</td>
</tr>
</tbody>
</table>

### Table 3: Initial Mapping Generated by Each Approach, Using Single-Word Names Only

<table>
<thead>
<tr>
<th>Approach</th>
<th>Coverage</th>
<th>Missed Connections</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular Expressions</td>
<td>82.1%</td>
<td>21/169</td>
<td>87.6%</td>
<td>56.7%</td>
<td>68.8%</td>
</tr>
<tr>
<td>Class List</td>
<td>41.7%</td>
<td>82/169</td>
<td>51.5%</td>
<td>87.9%</td>
<td>64.9%</td>
</tr>
<tr>
<td>Textwash</td>
<td>83.0%</td>
<td>24/169</td>
<td>85.8%</td>
<td>5.2%</td>
<td>9.7%</td>
</tr>
<tr>
<td>Regular Expressions + Class List</td>
<td>88.1%</td>
<td>16/169</td>
<td>90.5%</td>
<td>55.0%</td>
<td>68.5%</td>
</tr>
<tr>
<td>Textwash + Class List</td>
<td>84.5%</td>
<td>20/169</td>
<td>88.2%</td>
<td>5.3%</td>
<td>9.9%</td>
</tr>
</tbody>
</table>
Table 4: Counts of Unique Names, Connections, and Substitutions, by Category of Name

<table>
<thead>
<tr>
<th>Category of Name</th>
<th>Unique Names</th>
<th>Connections</th>
<th>Substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registered Name of Participant</td>
<td>83</td>
<td>88</td>
<td>2,169</td>
</tr>
<tr>
<td>Nickname, Shortened, or Stylised Form</td>
<td>36</td>
<td>41</td>
<td>669</td>
</tr>
<tr>
<td>Misspelled Name</td>
<td>34</td>
<td>34</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>148</td>
<td>163</td>
<td>2,888</td>
</tr>
</tbody>
</table>

There was also one example of an initial letter being used as a sign off; the mapping did not include the initial X, but it did contain the phrase Thanks X. The initial would be added to the mapping by the basic edits described in Section 5.3. Therefore, of the 28 missing connections, minimal manual intervention could be expected to restore 16 of them.

The remaining 12 missing connections all related to names that were not identified by the regular expressions and were not on the class list. This group of missing connections affected 11 individuals and accounted for 31 missed substitutions. Of these, 4 were nicknames, initials, or stylised forms of names. Another 6 missed names were one-off spelling errors, like Arhtur. For comparison, when the Textwash model was used to identify possible names, several of the misspelled names were again missed. Of the 34 misspelled names in the data set, the Textwash model found only 16.

5.5 Grouping Personal Names for Substitution

We found that substituting the names session-by-session worked best. There were 4 cases where duplicate names were found within a session; these were resolved manually. Taking each message thread separately risked missing mentions of the personal names of participants who did not post in that thread but were nevertheless known to the other participants. Replacing the names across the full data set at once generated 12 spurious duplicate name warnings.\(^6\) Figure 5 shows the output from the full Textwash system when it was used directly to anonymise the example messages. The Textwash system assumes all instances of the same name are references to the same individual, such as when the name Arthur appeared twice in the first example message (Figure 1). In fact, the middle initial C was also connected to the same individual, due to its use in the name Arthur C. Clarke, and would be substituted with that individual’s identifier even if it appeared alone in another message. Textwash has no concept of duplicate names and cannot generate warnings about them. There is also no way to indicate to Textwash that variants of a name refer to the same individual, such as Mary Jane, Mary, and MJ. Textwash has no mechanism to add missed names, such as Robert in the final example message.

---

\(^6\)For example, if two participants named Robert were enrolled on the course in different sessions, there could be no real ambiguity in any given message about which person was being addressed, so a warning about the duplicate name would be considered spurious.
6 DISCUSSION

Our exploration of personal names used in a data set of discussion forum messages reinforced the importance of taking a data-driven approach to name discovery. If anonymisation or pseudonymisation relied on the class list alone, a large fraction of the personal names would be missed, potentially revealing the identity of the participants and compromising their privacy [7]. The most common category of alternative names in our data was that of nicknames and shortened forms, such as Robbie or Bob instead of Robert. In many cases, such names could be used to identify an individual just as easily as a full name. Sharing or publishing such data would be unethical [7], potentially illegal [23], and would certainly constitute a breach of trust with the participants [14].

Misspelled names accounted for a much smaller proportion of the names in the data set, but were also more difficult to identify. The regular expressions were successful in identifying 26 of 34 misspelled names, while the BERT-based model used in Textwash found only 16. Future work could adapt the set-theoretic approach of Bosch and colleagues [1] to find unknown words, and then compare the edit-distance between each known name and the unknown words to identify misspellings like Arturr.

The approach outlined in this work is widely applicable to other areas of educational research that make use of informal written messages exchanged between participants, although the details of the name identification step will vary with the data. The metadata we used is commonly available: an identifier for the person who posted each message, an identifier for the message itself, and an identifier for the parent message (if any). With this small amount of information, the candidate names found in each message can be tentatively connected with the participants, even in the absence of a class list [25]. In contrast, a general-purpose pseudonymisation tool like Textwash does not provide any mechanism for incorporating such domain knowledge.

We proposed two essential requirements for a pseudonymisation tool in the field of learning analytics:

- the ability to connect together multiple names for the same individual, and
- the ability to track and resolve duplicate names.

Both of these requirements are necessary in order to ensure that the resulting data is useful for researchers who want to track the flow of ideas between conversation participants, a use case that is not supported by simply masking all the names [4, 12, 23]. We found that the majority of participants in our data set were referred to by more than one name (Figure 4). By connecting those names together, we gained a fuller picture of each participant’s input and interactions. Additionally, even a small data set may contain examples where more than one participant is referred to by the same name, leading to duplicate names in the data. These duplicates are often ignored [15, 19], but without the ability to identify and resolve such cases, the resulting pseudonymisation would be confusing and potentially misleading.

6.1 Limitations

The specific regular expressions we used to collect candidate names were simple but proved to be effective. In another data set, the pattern of changes between participants will be different, and different regular expressions (or a different approach entirely) would be required at step 1 to gather an initial list of words and phrases that might be personal names. In step 2, there could be regional variations in the forms of names and nicknames that are considered valid, beyond what was seen in our data. It might also prove worthwhile to collect a list of names that should not be replaced at step 3 – for example, names of public figures from Wikipedia, or authors of reference books relevant to the domain.

6.2 Conclusion

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Analytics for Social Impact and Physical Sciences Research Council (grant EP/L16427/1) and the University of Edinburgh.

ACKNOWLEDGMENTS

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Analytics for Social Impact and Physical Sciences Research Council (grant EP/L16427/1) and the University of Edinburgh.

REFERENCES


The edit distance is calculated as the number of character insertions and deletions needed to transform one word into another.


2.3 Summary of contributions

This chapter introduced the data used in this thesis and illustrated the structure of the nested discussion threads. Labels that were reused from earlier studies were described and data annotations specific to this programme of research were introduced and motivated.

We presented our novel data-driven approach for replacing the real names of the discussion participants with consistent pseudonyms. The primary motivation for replacing personal names was to allow independent annotators to label the messages without discovering the real identities of any of the discussion participants. A post hoc analysis of the pseudonymised corpus revealed that nearly one-quarter (24.9%) of all instances of personal names that were used to identify discussion participants were shortened forms, nicknames, or misspellings. We are not aware of any previous studies that quantified the prevalence of such names in discussion forum messages. Any approach to pseudonymisation that relied on the class list alone would miss names of this type, potentially revealing the identities of those participants.

Reconciling the individual participants with the various name forms used for them was important for the purposes of the research reported in this thesis because it allowed the flow of discussion to be tracked through each conversation thread. In particular, it allowed the annotators to correctly identify longer-distance references, such as from message $G_1$ back to message $D$ in Figure 2.1(b). In turn, identifying the connections between messages enabled the messages to be correctly labelled with the ICAP modes of cognitive engagement, supporting the studies in the next chapters.
Chapter 3

Markers of cognitive quality

3.1 Introduction

The focus of this chapter is on examining how linguistic and structural attributes of the dialogue that takes place in asynchronous online discussions are associated with the theorised properties of cognitive quality. Since cognitive quality is defined in terms of both the CoI phases of cognitive presence and the ICAP modes of cognitive engagement, we used the data set that was introduced in Chapter 2, where the messages were annotated with labels for each of those constructs. We looked for possible associations between the message labels and the linguistic and structural attributes of the messages. The work in this chapter thus addresses the first research question posed in Section 1.1, by assessing the extent to which the theorised properties of cognitive quality were associated with the dialogue attributes. Examining those associations also provided some initial insight into how the CoI and ICAP frameworks are related to each other.

While the primary goal of discussion forum participation is to support students’ learning (Gašević, Dawson, et al., 2015; Wang et al., 2015; Wise & Cui, 2018), the content of the discussions can also be useful for researchers (Garrison et al., 2001; Wise et al., 2016). The messages that participants exchange vary in quality (Corich et al., 2004), and those differences may be associated with attributes of the dialogue, such as sentence length, position within the discussion thread, and word choice.

When students interact in discussion forums that are linked to a course of study, their messages often contain specialised, domain-specific, terminology that is related to the course topic. For example, in the case of a software engineering course, technical terms might be used, such as “middleware” and “hardware virtualization”. Domain-
specific words such as these are unlikely to be directly informative for understanding the cognitive quality of a contribution, and will not generalise to courses in different domains. However, other attributes of the dialogue may be usefully correlated with quality, either positively or negatively. Linguistic attributes of the dialogue, such as whether messages are framed as questions, whether they address other participants directly using second-person pronouns, and whether they make reference to the content of other participants’ contributions, have all proven to be useful in earlier studies of cognitive presence (Kovanović et al., 2016; Neto et al., 2018; Barbosa et al., 2020).

The location of a message within the structure of the dialogue can also be informative (Rosé et al., 2008; Anderson & Dron, 2011; Wise et al., 2014; Waters et al., 2015; Kovanović et al., 2016). Messages posted deep within a thread are generally expected to relate to what has been said in messages located above them in the thread. For example, they may summarise or critique earlier messages. In contrast, messages that belong to different branches of the discussion may not take into account the content included in a neighbouring branch (Wise et al., 2014).

Many dialogue attributes can be calculated directly from the message content, including the number of words and sentences, and the prevalence of certain classes of words, such as prepositions. Other metrics, such as readability scores, have been calibrated based on empirical studies (McNamara et al., 2014, p. 80). The linguistic dialogue attributes we used were word counts and coherence metrics derived using the LIWC and Coh-Metrix software packages. These were the same attributes as had been used in earlier studies, allowing our results to be compared directly with those studies. In addition, we derived several structural attributes for each message. We measured the chronological position of the message from the beginning and from the end of the thread, and calculated the depth of the message within the threaded reply structure. We also counted both the number of direct replies to each message and the total number of nested replies.¹

The work described in this chapter considered dialogue attributes, both linguistic and structural, as potential markers of cognitive quality. In Chapter 1, we defined cognitive quality as students’ intellectual engagement with the educational content of the discussion, as measured by both the CoI phases of cognitive presence and the ICAP modes of cognitive engagement. In the publication that is included in the next section, we first examined the predictive power of the dialogue attributes in relation to

¹For example, message C in Figure 2.1(a) has a depth of 1 and a chronological position of 6 from the start of the thread, and of 2 from the end. The message that started the thread in the same example received 3 direct replies and 6 replies in total.
the CoI phases of cognitive presence and the ICAP modes of cognitive engagement, considered separately. Then, in order to gain some initial insight into how the two frameworks relate to one another, we went on to use the labels from each framework as additional model features alongside the dialogue attributes.

### 3.2 Peer-reviewed publication: Markers of Cognitive Quality in Student Contributions to Online Course Discussion Forums

This section includes the verbatim copy of the following peer-reviewed publication, reprinted with permission:


**Contributions:** The ideas and analysis in the paper were developed and discussed between all authors of the work. The original idea, the experiments, and the writing were the work of the first author.

Supplementary material from the publication is included as Appendix B.
Markers of Cognitive Quality in Student Contributions to Online Course Discussion Forums

Elaine Farrow¹, Johanna D. Moore², Dragan Gašević³

Abstract
By participating 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, we also assessed the links between frameworks. 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 analyze 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 than using either one alone.

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

Submitted: 05/06/2020 — Accepted: 18/03/2022 — Published: 31/08/2022

Corresponding author ¹ Email: elaine.farrow@ed.ac.uk Address: School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, UK. ORCID ID: https://orcid.org/0000-0002-1152-3443
² Email: j.moore@ed.ac.uk Address: School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, UK. ORCID ID: https://orcid.org/0000-0001-7247-6823
³ Email: dragan.gasevic@monash.edu Address: Faculty of Information Technology, Monash University, Clayton 3800, Australia. ORCID ID: https://orcid.org/0000-0001-9265-1908

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)
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 just to clarify their own thoughts. Earlier work showed that learning can take place by 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; Wang et al., 2016; Wise & Cui, 2018) shows 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 for the number of messages generated in a discussion forum to be 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; Wang et al., 2016; Kovanović et al., 2016; R. Ferreira et al., 2018; Neto et al., 2018; Farrow et al., 2019; Yoge 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, instructors could identify students who are bored, frustrated, or struggling, or lessons that cause confusion, while there is still time to intervene. For example, discussion forum transcripts could be colour-coded to indicate how the conversation is progressing, enabling instructors to see at a glance where to direct their attention (Yoge 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 (CoI) 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 these 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 CoI 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 important, 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 (Stevenson et al., 2017), but these were too distant from our primary focus. Stories 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 of 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 and 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—they 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 these 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 framework.
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 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 CoI Framework

The CoI framework for online education is a powerful tool for analyzing 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:

1. a social environment conducive to learning (social presence),
2. a well-designed course with ongoing facilitation (teaching presence), and
3. 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
Figure 1. An Idealized Example of How a Discussion Might Progress through the CoI Phases of Cognitive Presence

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 ω = 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.

Col has been widely used to analyze student learning in online courses, and predictive models have been developed to identify its elements automatically using the text of discussion forum messages (McKlin, 2004; Corich et al., 2006; 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 and colleagues (2015) investigated the effect of varying the scaffolding provided to different cohorts of students, an aspect of teaching presence. Recent work by M. Ferreira and colleagues (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 ω = 0.70). Inputs were mainly dictionary-based features, along with features describing the position of the message in the threaded discussion. Corich and colleagues (2006) labelled the sentences within messages, rather than whole messages, using an automated content analysis tool. Kovanović and colleagues (2014) used support vector machines to label messages with Col phases of cognitive presence. They used standard bag-of-words text features without adding any features to account for context. Waters and colleagues (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 analyzed 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; Yoge et al., 2018; Farrow et al., 2019; Barbosa et al., 2020; M. Ferreira et al., 2020). Yoge and colleagues (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 and colleagues (2016) used a random forest together with both structural and high-level linguistic features. It claimed high accuracy (Cohen’s ω = 0.63), but a later replication study (Farrow et al., 2019) indicated that the model’s predictive power was likely to have been overestimated (Cohen’s ω = 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 ω = 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 ω = 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 interactive, constructive, active, and passive. 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.

![Hierarchical Modes of Cognitive Engagement in the ICAP Framework](image)

**Figure 2.** The Hierarchical Modes of Cognitive Engagement in the ICAP Framework

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 observed mode of engagement. 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 pre-recorded video, then the activity is 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; Wang et al., 2016) and to student comments on an annotated electronic course text (Yogev et al., 2018) and on MOOC videos (Taskin et al., 2019). Atapattu and colleagues (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 three 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:

- **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

**Constructive extending:**
- elaborating on a point or displaying reasoning about course content;
- proposing ideas, sharing resources, or asking questions going beyond course materials;

**Active targeted:** referring explicitly to course content by paraphrasing or asking clarification questions;
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 posted only 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 in the group with interactive and/or constructive messages had a larger effect size than membership in the active group.

The ICAP framework has also been used to analyze 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 and colleagues (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 (i.e., 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 were 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 (i.e. & 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; Wang et al., 2016), very few posts were labelled as interactive, while in the video annotation study (Taskin et al., 2019), interactive posts 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 emphasize 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 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 (Schröre, 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 apply to work using the other. For example, learning interventions that encourage 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 target 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 idealized 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 idealized 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 is therefore as follows:

**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-level 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 is as follows:

- **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 are expected to be similar. When a random forest is constructed, only a subset of the available features is used for each tree in the forest. Therefore, in some trees the most highly predictive features are not used, which 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 uses a data set that has previously been used in several studies of cognitive presence (Kovanović et al., 2014; Galevič 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 Master’s 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.

Each student created and shared a video presentation based on a research paper relevant to the course and then started a new thread in the discussion forum to host a conversation about their presentation. Every student was expected to lead at least one discussion thread and contribute to at least three others. We do not have access to the presentations themselves, only to the text-based discussions that followed. Students were in general highly motivated, since the first assignment accounted for 10%
Table 1. Statistics for the Six Course Offerings Used in This Work

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

of the final course mark. In contrast to the MOOC data used in Wang and colleagues (2016) and Taskin and colleagues (2019), genuine interaction between students was common 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 Phases of Cognitive Presence

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

The other class was used for messages that displayed no sign of cognitive presence.

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; Wang et al., 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 and colleagues (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 and colleagues (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 and colleagues (2019) considered them to indicate “shallow” engagement. Hu and colleagues (2020) distinguished between agreements that were justified with reasoning and those that stood alone; the former were considered to demonstrate deeper cognitive presence than the latter. Yogev and colleagues (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.
3.2. Peer-reviewed publication: Markers of Cognitive Quality

The extended taxonomy was adapted from Yogev and colleagues (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.

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 five discussion threads and resolved disagreements through discussion, following which the annotation guidelines were clarified further. A further 12 discussion threads (202 messages) were labelled independently, and the labels from these were used to assess inter-annotator agreement at 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 using time-stamp order and then derived features using that ordering (position from start, position from end, and fractional 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 descendent 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

---

**Table 3. Breakdown of Messages across Modes in the Extended Cognitive Engagement Taxonomy**

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

The revised annotation guidelines are available from https://homepages.inf.ed.ac.uk/efarrow/Resources.
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 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 in the threaded or chronological order.

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.

### 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). Since there were so few off-task messages, those records were excluded...
Table 4. Structural Attributes Derived from the Data and Used as Model Features

<table>
<thead>
<tr>
<th>Structural attribute</th>
<th>Feature Description</th>
<th>rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth in thread</td>
<td>message.depth</td>
<td>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.</td>
</tr>
<tr>
<td>First/last message</td>
<td>message.is.first, message.is.last</td>
<td>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.</td>
</tr>
<tr>
<td>Number of direct replies</td>
<td>message.replies.direct</td>
<td>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).</td>
</tr>
<tr>
<td>Total number of replies</td>
<td>message.replies.all</td>
<td>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).</td>
</tr>
<tr>
<td>Position from start</td>
<td>message.pos.start</td>
<td>The index of the message in chronological order from the beginning of the discussion. Early messages may be more likely to introduce new material.</td>
</tr>
<tr>
<td>Position from end</td>
<td>message.pos.end</td>
<td>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.</td>
</tr>
<tr>
<td>Fractional position</td>
<td>message.pos.frac</td>
<td>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.</td>
</tr>
<tr>
<td>Discussion size</td>
<td>message.thread.size</td>
<td>The total number of messages in the current discussion. We hypothesize that a short discussion is less likely to progress to deeper phases of cognitive presence than a longer one.</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Cognitive presence phase</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Percentage</td>
</tr>
<tr>
<td>Triggering event</td>
<td>280</td>
<td>18.54%</td>
</tr>
<tr>
<td>Exploration</td>
<td>608</td>
<td>40.26%</td>
</tr>
<tr>
<td>Integration</td>
<td>425</td>
<td>28.15%</td>
</tr>
<tr>
<td>Resolution</td>
<td>85</td>
<td>5.63%</td>
</tr>
<tr>
<td>Other</td>
<td>112</td>
<td>7.42%</td>
</tr>
<tr>
<td>All</td>
<td>1510</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

We explored 20 different settings for the mtry parameter, which 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 practices, the SMOTE algorithm was run inside the

---

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)
Table 6. Messages by ICAP Modes of Cognitive Engagement in Training and Test Partitions

<table>
<thead>
<tr>
<th>Cognitive engagement mode</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Percentage</td>
</tr>
<tr>
<td>Active</td>
<td>313</td>
<td>20.73%</td>
</tr>
<tr>
<td>Affirmation</td>
<td>66</td>
<td>4.37%</td>
</tr>
<tr>
<td>Constructive</td>
<td>616</td>
<td>40.79%</td>
</tr>
<tr>
<td>Interactive</td>
<td>506</td>
<td>33.51%</td>
</tr>
<tr>
<td>Off-task</td>
<td>9</td>
<td>0.60%</td>
</tr>
<tr>
<td>All</td>
<td>1510</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

The messages labelled as *off-task* were excluded from the analysis.

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.

Table 7. Models Trained in Experiment 1 for Each of the Two Frameworks, with and without Class Rebalancing

<table>
<thead>
<tr>
<th>Model</th>
<th>Outcome classes</th>
<th>LIWC</th>
<th>Coh-Metrix</th>
<th>Structural</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoI phases of cognitive presence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model CoI-1a</td>
<td>Five phases, original distribution</td>
<td>91</td>
<td>106</td>
<td>9</td>
<td>206</td>
</tr>
<tr>
<td>Model CoI-1b</td>
<td>Five phases, rebalanced with SMOTE</td>
<td>91</td>
<td>106</td>
<td>9</td>
<td>206</td>
</tr>
<tr>
<td>ICAP modes of cognitive engagement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model ICAP-1a</td>
<td>Four modes, original distribution</td>
<td>91</td>
<td>106</td>
<td>9</td>
<td>206</td>
</tr>
<tr>
<td>Model ICAP-1b</td>
<td>Four modes, rebalanced with SMOTE</td>
<td>91</td>
<td>106</td>
<td>9</td>
<td>206</td>
</tr>
</tbody>
</table>

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. 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 five 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 four ICAP modes, while Model ICAP-6 used those same labels to predict the extended set of six ICAP modes. These models and their features are summarized 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 one, while the features corresponding to the other ICAP labels were all assigned the value zero. 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 one, while the features corresponding to the other CoI labels would all have the value zero.

<table>
<thead>
<tr>
<th>Table 8. Models Trained in Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>CoI phases of cognitive presence</td>
</tr>
<tr>
<td>Model CoI-2a</td>
</tr>
<tr>
<td>Model CoI-2b</td>
</tr>
<tr>
<td>ICAP modes of cognitive engagement</td>
</tr>
<tr>
<td>Model ICAP-4</td>
</tr>
<tr>
<td>Model ICAP-6</td>
</tr>
</tbody>
</table>

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

We prepared the data as before, using the first five offerings of the course as training data and 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 \( \kappa \) and the macro-averaged \( F_1 \) score are more informative than accuracy. We chose the best model for each framework based on Cohen’s \( \kappa \) (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>
<thead>
<tr>
<th>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</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Col phases of cognitive presence</td>
</tr>
<tr>
<td>Model Col-1a</td>
</tr>
<tr>
<td>Model Col-1b</td>
</tr>
<tr>
<td>ICAP modes of cognitive engagement</td>
</tr>
<tr>
<td>Model ICAP-1a</td>
</tr>
<tr>
<td>Model ICAP-1b</td>
</tr>
</tbody>
</table>

The best results for each framework are in bold.

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 \( \kappa \) 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 Col phases of cognitive presence, a Cohen’s \( \kappa \) of 0.358 indicates a “fair” level of agreement with the gold-standard human coding, while the Cohen’s \( \kappa \) 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 Col-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 its accounting for only 4.37% of the training data.

### Table 10. Experiment 1: Outcome Metrics on the Held-Out Test Data

<table>
<thead>
<tr>
<th>Outcome classes</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Macro $F_1$</th>
<th>Cohen’s $κ$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Col phases of cognitive presence, using Model Col-1b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triggering event</td>
<td>0.719</td>
<td>0.821</td>
<td>0.767</td>
<td>0.515</td>
<td>0.358</td>
</tr>
<tr>
<td>Exploration</td>
<td>0.460</td>
<td>0.526</td>
<td>0.491</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integration</td>
<td>0.543</td>
<td>0.530</td>
<td>0.537</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution</td>
<td>0.250</td>
<td>0.136</td>
<td>0.176</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0.640</td>
<td>0.571</td>
<td>0.604</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICAP modes of cognitive engagement, using Model ICAP-1b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td>0.782</td>
<td>0.681</td>
<td>0.719</td>
<td></td>
<td>0.764</td>
</tr>
<tr>
<td>Affirmation</td>
<td>0.571</td>
<td>0.889</td>
<td>0.696</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constructive</td>
<td>0.857</td>
<td>0.793</td>
<td>0.824</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactive</td>
<td>0.778</td>
<td>0.863</td>
<td>0.818</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 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 MDG values (Figure 5). The top 20 features by importance for each framework are listed in Tables 11 and 12. Their distributions across the Col 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.

**Figure 5.** MDG 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.DESMC) appears in the top five for both models. Similarly, lower levels of lexical diversity, measured by type-token ratio (cm.LDTRa and cm.LDTRc), 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 deeper cognitive engagement. These results are in line with prior work on Col (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 strongly indicate other messages (those that display no signs of cognitive presence) and 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 phase of cognitive presence and to indicate the interactive or affirmation mode of ICAP (Table 13). None of the features relating to time-stamp
## 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 MDG

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature Description</th>
<th>MDG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cm.DESWC</td>
<td>164.76</td>
</tr>
<tr>
<td>2</td>
<td>message.is.first</td>
<td>87.54</td>
</tr>
<tr>
<td>3</td>
<td>liwc.posemo</td>
<td>83.11</td>
</tr>
<tr>
<td>4</td>
<td>cm.WRDMEAc</td>
<td>75.55</td>
</tr>
<tr>
<td>5</td>
<td>message.depth</td>
<td>74.90</td>
</tr>
<tr>
<td>6</td>
<td>cm.LDTTRa</td>
<td>69.15</td>
</tr>
<tr>
<td>7</td>
<td>liwc.GemoC</td>
<td>65.84</td>
</tr>
<tr>
<td>8</td>
<td>liwc.OMark</td>
<td>57.07</td>
</tr>
<tr>
<td>9</td>
<td>cm.WRDHPn</td>
<td>48.46</td>
</tr>
<tr>
<td>10</td>
<td>message.replies.direct</td>
<td>47.84</td>
</tr>
<tr>
<td>11</td>
<td>liwc.affect</td>
<td>47.05</td>
</tr>
<tr>
<td>12</td>
<td>liwc.discrep</td>
<td>46.98</td>
</tr>
<tr>
<td>13</td>
<td>liwc.money</td>
<td>37.18</td>
</tr>
<tr>
<td>14</td>
<td>message.thread.size</td>
<td>34.78</td>
</tr>
<tr>
<td>15</td>
<td>message.replies.all</td>
<td>33.64</td>
</tr>
<tr>
<td>16</td>
<td>cm.DESWLtd</td>
<td>26.71</td>
</tr>
<tr>
<td>17</td>
<td>cm.DESWLtd</td>
<td>26.10</td>
</tr>
<tr>
<td>18</td>
<td>cm.LDVOCD</td>
<td>25.64</td>
</tr>
<tr>
<td>19</td>
<td>cm.LDTRc</td>
<td>22.89</td>
</tr>
<tr>
<td>20</td>
<td>liwc.near</td>
<td>21.47</td>
</tr>
</tbody>
</table>

Note: Features that also appear in the top 20 for the best ICAP model (Table 12) are shown in italics.

## 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 MDG

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature Description</th>
<th>MDG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>message.depth</td>
<td>145.94</td>
</tr>
<tr>
<td>2</td>
<td>liwc.assent</td>
<td>91.08</td>
</tr>
<tr>
<td>3</td>
<td>message.replies.direct</td>
<td>70.39</td>
</tr>
<tr>
<td>4</td>
<td>message.replies.all</td>
<td>62.65</td>
</tr>
<tr>
<td>5</td>
<td>cm.DESWC</td>
<td>61.24</td>
</tr>
<tr>
<td>6</td>
<td>cm.LDVOCD</td>
<td>59.97</td>
</tr>
<tr>
<td>7</td>
<td>cm.LDTRa</td>
<td>49.32</td>
</tr>
<tr>
<td>8</td>
<td>liwc.posemo</td>
<td>41.61</td>
</tr>
<tr>
<td>9</td>
<td>cm.LSAGH</td>
<td>39.99</td>
</tr>
<tr>
<td>10</td>
<td>cm.DESFL</td>
<td>36.66</td>
</tr>
<tr>
<td>11</td>
<td>cm.DESSC</td>
<td>36.21</td>
</tr>
<tr>
<td>12</td>
<td>liwc.OMark</td>
<td>35.95</td>
</tr>
<tr>
<td>13</td>
<td>liwc.affect</td>
<td>31.59</td>
</tr>
<tr>
<td>14</td>
<td>cm.LDTRc</td>
<td>22.39</td>
</tr>
<tr>
<td>15</td>
<td>liwc.Period</td>
<td>21.22</td>
</tr>
<tr>
<td>16</td>
<td>cm.RDFKGL</td>
<td>20.60</td>
</tr>
<tr>
<td>17</td>
<td>liwc.tentat</td>
<td>17.88</td>
</tr>
<tr>
<td>18</td>
<td>cm.RDFRE</td>
<td>17.86</td>
</tr>
<tr>
<td>19</td>
<td>cm.DESWLtd</td>
<td>17.42</td>
</tr>
<tr>
<td>20</td>
<td>liwc.ppron</td>
<td>17.13</td>
</tr>
</tbody>
</table>

Note: Features that also appear in the top 20 for the best CoI model (Table 11) are shown in italics.
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 strongly indicates triggering events (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.DESAGN) 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) indicates active mode, where quoting is expected, and affirmation messages.

Overall, the most predictive lexical features have similar MDG scores to 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.DESFL 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.

### 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 in Experiment 1, we trained models using 10-fold cross-validation and compared the Cohen’s $\kappa$ scores for the purpose of model selection (Table 14). Comparing the cross-validation results for Model Col-2a and Model Col-2b against Model Col-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 ($\kappa = 0.460$), 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Outcome classes</th>
<th>Best mtry</th>
<th>Macro $F_1$</th>
<th>Cohen’s $\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoI phases of cognitive presence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model CoI-2a</td>
<td>Five phases, rebalanced with SMOTE</td>
<td>35</td>
<td>0.553</td>
<td>0.460</td>
</tr>
<tr>
<td>Model CoI-2b</td>
<td>Five phases, rebalanced with SMOTE</td>
<td>57</td>
<td>0.551</td>
<td>0.459</td>
</tr>
<tr>
<td>ICAP modes of cognitive engagement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model ICAP-4</td>
<td>Four modes, rebalanced with SMOTE</td>
<td>35</td>
<td>0.710</td>
<td>0.657</td>
</tr>
<tr>
<td>Model ICAP-6</td>
<td>Six extended modes, rebalanced with SMOTE</td>
<td>57</td>
<td>0.587</td>
<td>0.590</td>
</tr>
</tbody>
</table>

The best results for each framework are in bold.

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 generalizability 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.

Table 15. Experiment 2: Outcome Metrics on the Held-Out Test Data

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Macro $F_1$</th>
<th>Cohen’s $\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoI phases of cognitive presence, using Model CoI-2a</td>
<td>0.694</td>
<td>0.893</td>
<td>0.781</td>
<td></td>
<td>0.546</td>
</tr>
<tr>
<td>Triggering event</td>
<td>0.488</td>
<td>0.553</td>
<td>0.519</td>
<td></td>
<td>0.404</td>
</tr>
<tr>
<td>Exploration</td>
<td>0.561</td>
<td>0.554</td>
<td>0.558</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integration</td>
<td>0.429</td>
<td>0.136</td>
<td>0.207</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution</td>
<td>0.692</td>
<td>0.643</td>
<td>0.667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td>0.546</td>
<td>0.404</td>
</tr>
<tr>
<td>ICAP modes of cognitive engagement, using Model ICAP-4</td>
<td>0.839</td>
<td>0.553</td>
<td>0.667</td>
<td></td>
<td>0.765</td>
</tr>
<tr>
<td>Active</td>
<td>0.615</td>
<td>0.889</td>
<td>0.737</td>
<td></td>
<td>0.695</td>
</tr>
<tr>
<td>Affirmation</td>
<td>0.838</td>
<td>0.830</td>
<td>0.834</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constructive</td>
<td>0.767</td>
<td>0.904</td>
<td>0.830</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Meanwhile, Model ICAP-4 obtained “substantial” agreement (Cohen’s $\kappa = 0.695$) with the gold-standard labels for the 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, leaving the overall macro-averaged $F_1$ and Cohen’s $\kappa$ 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 MDG values across the model features is 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.

Mean Decrease Gini

Feature
0 20 40 60 80 100 150 200

(a) Model Col-2a

(b) Model ICAP-4

Figure 6. MDG 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 appear in the top 20 as in Experiment 1, in approximately the same order. The one-hot attributes generated from the framework constructs are 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 Col-2a, which labels the CoI phases of cognitive presence (Table 16). The discriminative power of the ICAP-1 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 indicates the other class. ICAP-A comes 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 are lower than for the other classes, although the standard deviation is relatively high across all classes. In contrast, ICAP-G, indicating off-task messages, appears 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 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.

In Model ICAP-4, three of the five binary attributes derived from the CoI phases of cognitive presence are 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 find CoI-other, strongly indicating affirmation messages. CoI-exploration, at position 15, has a smoother distribution and was most likely to indicate a constructive message. The attributes indicating integration and resolution do not appear in the top 20 list: CoI-integration is number 29, while CoI-resolution is number 196 out of 211. Sixteen of the remaining top 20 attributes also appear in the top 20 list in Experiment 1. The standard deviation of the mean number of syllables (cm. DESW Leyd) is a new addition at position 20; the highest values are likely to indicate constructive mode.

We have already seen that adding model features derived from the CoI phases of cognitive presence does not improve the overall predictive power of the model. The MDG values allow us to see how much information each of these features provides 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, and then we 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 Col

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 Col-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.
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 MDG

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
<th>Description</th>
<th>Previous rank</th>
<th>MDG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cm.DESWC</td>
<td>Number of words</td>
<td>1</td>
<td>110.76</td>
</tr>
<tr>
<td>2</td>
<td>ICAP-I</td>
<td>Labelled as interactive</td>
<td>–</td>
<td>95.25</td>
</tr>
<tr>
<td>3</td>
<td>ICAP-F</td>
<td>Labelled as affirmation</td>
<td>–</td>
<td>87.71</td>
</tr>
<tr>
<td>4</td>
<td>message.is.first</td>
<td>First message</td>
<td>2</td>
<td>64.76</td>
</tr>
<tr>
<td>5</td>
<td>message.depth</td>
<td>Message depth in discussion</td>
<td>5</td>
<td>62.81</td>
</tr>
<tr>
<td>6</td>
<td>ICAP-A</td>
<td>Labelled as active</td>
<td>–</td>
<td>55.19</td>
</tr>
<tr>
<td>7</td>
<td>cm.LDTRa</td>
<td>Lexical diversity, all words</td>
<td>6</td>
<td>51.11</td>
</tr>
<tr>
<td>8</td>
<td>liwc.SemiC</td>
<td>Number of semicolons</td>
<td>7</td>
<td>50.66</td>
</tr>
<tr>
<td>9</td>
<td>liwc.pos emo</td>
<td>Number of +ve emotion words</td>
<td>3</td>
<td>50.59</td>
</tr>
<tr>
<td>10</td>
<td>cm.WRDMEAc</td>
<td>Meaningfulness</td>
<td>4</td>
<td>48.53</td>
</tr>
<tr>
<td>11</td>
<td>message.replies.direct</td>
<td>Number of direct replies</td>
<td>10</td>
<td>45.89</td>
</tr>
<tr>
<td>12</td>
<td>cm.WRDRypn</td>
<td>Hypernyms for nouns</td>
<td>9</td>
<td>37.22</td>
</tr>
<tr>
<td>13</td>
<td>liwc.QMark</td>
<td>Number of question marks</td>
<td>8</td>
<td>36.85</td>
</tr>
<tr>
<td>14</td>
<td>message.replies.all</td>
<td>Total number of replies</td>
<td>15</td>
<td>36.76</td>
</tr>
<tr>
<td>15</td>
<td>liwc.discrep</td>
<td>Number of discrepancy words</td>
<td>12</td>
<td>36.45</td>
</tr>
<tr>
<td>16</td>
<td>liwc.affect</td>
<td>Number of affective process words</td>
<td>11</td>
<td>32.38</td>
</tr>
<tr>
<td>17</td>
<td>liwc.money</td>
<td>Number of money words</td>
<td>13</td>
<td>29.20</td>
</tr>
<tr>
<td>18</td>
<td>message.thread.size</td>
<td>Discussion size</td>
<td>14</td>
<td>26.63</td>
</tr>
<tr>
<td>19</td>
<td>cm.DESWLTld</td>
<td>SD of word length in letters</td>
<td>17</td>
<td>25.12</td>
</tr>
<tr>
<td>20</td>
<td>ICAP-C</td>
<td>Labelled as constructive</td>
<td>–</td>
<td>23.21</td>
</tr>
</tbody>
</table>

Note: 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 in italics.

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 MDG

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
<th>Description</th>
<th>Previous rank</th>
<th>MDG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>message.depth</td>
<td>Message depth in discussion</td>
<td>1</td>
<td>150.99</td>
</tr>
<tr>
<td>2</td>
<td>CoI-triggering</td>
<td>Labelled as triggering event</td>
<td>–</td>
<td>104.89</td>
</tr>
<tr>
<td>3</td>
<td>liwc.assent</td>
<td>Number of expressions of assent</td>
<td>2</td>
<td>80.78</td>
</tr>
<tr>
<td>4</td>
<td>message.replies.direct</td>
<td>Number of direct replies</td>
<td>3</td>
<td>70.14</td>
</tr>
<tr>
<td>5</td>
<td>message.replies.all</td>
<td>Total number of replies</td>
<td>4</td>
<td>52.77</td>
</tr>
<tr>
<td>6</td>
<td>cm.DESWC</td>
<td>Number of words</td>
<td>5</td>
<td>52.12</td>
</tr>
<tr>
<td>7</td>
<td>cm.LDVOCD</td>
<td>Lexical diversity, VOCD</td>
<td>6</td>
<td>43.53</td>
</tr>
<tr>
<td>8</td>
<td>cm.LSAOB</td>
<td>LSA given-new ratio</td>
<td>9</td>
<td>42.24</td>
</tr>
<tr>
<td>9</td>
<td>cm.LDTRa</td>
<td>Lexical diversity, all words</td>
<td>7</td>
<td>40.19</td>
</tr>
<tr>
<td>10</td>
<td>CoI-other</td>
<td>Labelled as other</td>
<td>–</td>
<td>38.23</td>
</tr>
<tr>
<td>11</td>
<td>liwc.pos emo</td>
<td>Number of +ve emotion words</td>
<td>8</td>
<td>36.03</td>
</tr>
<tr>
<td>12</td>
<td>cm.DESPL</td>
<td>Mean length of paragraphs</td>
<td>10</td>
<td>33.62</td>
</tr>
<tr>
<td>13</td>
<td>cm.DESBC</td>
<td>Number of sentences</td>
<td>11</td>
<td>33.01</td>
</tr>
<tr>
<td>14</td>
<td>liwc.affect</td>
<td>Number of affective process words</td>
<td>13</td>
<td>26.58</td>
</tr>
<tr>
<td>15</td>
<td>CoI-exploration</td>
<td>Labelled as exploration</td>
<td>–</td>
<td>26.29</td>
</tr>
<tr>
<td>16</td>
<td>liwc.QMark</td>
<td>Number of question marks</td>
<td>12</td>
<td>25.17</td>
</tr>
<tr>
<td>17</td>
<td>cm.LDTRa</td>
<td>Lexical diversity, content words</td>
<td>14</td>
<td>19.22</td>
</tr>
<tr>
<td>18</td>
<td>cm.DESWLTld</td>
<td>SD of word length in letters</td>
<td>19</td>
<td>17.26</td>
</tr>
<tr>
<td>19</td>
<td>liwc.Period</td>
<td>Number of periods</td>
<td>15</td>
<td>17.13</td>
</tr>
<tr>
<td>20</td>
<td>cm.DESWLSyd</td>
<td>SD of the mean number of syllables</td>
<td>–</td>
<td>15.43</td>
</tr>
</tbody>
</table>

Note: 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 in italics.
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 $\kappa = 0.623$). However, our experiments used only four ICAP modes, rather than the six 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 six 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 five 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) emphasizes 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 that 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 useful context 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 hypothesized 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

---

### 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

<table>
<thead>
<tr>
<th>ICAP</th>
<th>CoI-triggering</th>
<th>CoI-exploration</th>
<th>CoI-integration</th>
<th>CoI-resolution</th>
<th>CoI-other</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICAP-A</td>
<td>0.53 (0.50)</td>
<td>0.00 (0.00)</td>
<td>0.47 (0.50)</td>
<td>0.01 (0.08)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>ICAP-F</td>
<td>0.19 (0.39)</td>
<td>0.03 (0.17)</td>
<td>0.48 (0.50)</td>
<td>0.30 (0.46)</td>
<td>0.00 (0.04)</td>
</tr>
<tr>
<td>ICAP-C</td>
<td>0.02 (0.15)</td>
<td>0.01 (0.12)</td>
<td>0.41 (0.49)</td>
<td>0.56 (0.50)</td>
<td>0.00 (0.04)</td>
</tr>
<tr>
<td>ICAP-I</td>
<td>0.01 (0.10)</td>
<td>0.01 (0.10)</td>
<td>0.22 (0.42)</td>
<td>0.76 (0.43)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>ICAP-O</td>
<td>0.41 (0.49)</td>
<td>0.34 (0.47)</td>
<td>0.14 (0.35)</td>
<td>0.04 (0.20)</td>
<td>0.06 (0.25)</td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>CoI</th>
<th>ICAP-A</th>
<th>ICAP-F</th>
<th>ICAP-C</th>
<th>ICAP-I</th>
<th>ICAP-O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>0.45 (0.50)</td>
<td>0.35 (0.48)</td>
<td>0.03 (0.18)</td>
<td>0.00 (0.05)</td>
<td>0.16 (0.37)</td>
</tr>
<tr>
<td>Affirmation</td>
<td>0.00 (0.00)</td>
<td>0.27 (0.45)</td>
<td>0.09 (0.29)</td>
<td>0.01 (0.12)</td>
<td>0.63 (0.49)</td>
</tr>
<tr>
<td>Constructive</td>
<td>0.20 (0.40)</td>
<td>0.45 (0.50)</td>
<td>0.29 (0.45)</td>
<td>0.03 (0.18)</td>
<td>0.03 (0.16)</td>
</tr>
<tr>
<td>Interactive</td>
<td>0.00 (0.06)</td>
<td>0.36 (0.48)</td>
<td>0.49 (0.50)</td>
<td>0.14 (0.35)</td>
<td>0.01 (0.10)</td>
</tr>
</tbody>
</table>
the individual phases, with macro-averaged $F_1$ scores improving from 0.515 to 0.546 and Cohen’s $\kappa$ 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 were direct analogues 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 $\kappa$ 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 events 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 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 (Yoge 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 so 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 these 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 to that of 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 and colleagues (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. The study by Kovanović and colleagues (2016) 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. These 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 that 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. The features used in the current work and related previous studies are summarized 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...
Comparing the top 20 most predictive features from Experiment 1 (Table 11) with those in the top 20 reported by Kovanović and colleagues (2016), eight of a possible 18 common features appeared in both lists. The replication study (Farlow et al., 2019) did not report which features were most predictive in the revised model. In fact, only nine 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 high in the current study may not have been available to those classifiers (e.g., cm. \text{WRDMEAc}, defined as “meaningfulness”). Our best estimate is that six of a possible 12 common features from Experiment 1 are also in the top 20 for the Portuguese-only study (Neto et al., 2018), while three of 11 common features appear in the top 20 for the cross-language study (Barbosa et al., 2020). The comparative rankings of features are summarized 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 (Yoge 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 CoL phases of cognitive presence.

5.7 Implications for Research and Practice
Conceptual frameworks like CoL 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; Farlow 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 overview of the discussion quality while there is time to intervene (Yoge 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 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 high 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 required (Vellikkunel et al., 2017). To overcome this problem, we suggest that educators 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 CoL 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

ISSN 1939-7750 (online), The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)
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 prioritize 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 (Yoge 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 than simply clicking a symbol; it also provides scope for personalization 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 hypothesize 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 indicate that the frameworks are not closely aligned and suggest 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; Farow et al., 2019; M. Ferreira et al., 2020; Hu et al., 2020) and ICAP (Yoge 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 authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
3.2. Peer-reviewed publication: Markers of Cognitive Quality

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


ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution · NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)
3.2. Peer-reviewed publication: Markers of Cognitive Quality


3.3 Summary of contributions

In this chapter, we examined possible associations between the linguistic and structural attributes of messages collected from an asynchronous online discussion, and the theorised properties of cognitive quality. The two sets of constructs that together define cognitive quality, the CoI phases of cognitive presence and the ICAP modes of cognitive engagement, were treated separately. Results from earlier work, looking at dialogue attributes that were predictive of specific CoI phases of cognitive presence, were confirmed. The ICAP modes of cognitive engagement had never previously been examined in connection with dialogue attributes. Our study identified some dialogue attributes that were important for predicting both the ICAP modes of cognitive engagement and the CoI phases of cognitive presence; these attributes can thus be considered to be predictive for cognitive quality (Table 3.1). Overall, the results indicated that the two frameworks were measuring different aspects of cognitive quality. Later chapters of this thesis go on to use the CoI phases of cognitive presence and the ICAP modes of cognitive engagement together to present a two-dimensional picture of cognitive quality.

One notable finding of the study presented in this chapter related to the distinction between the order in which contributions are added to a discussion, compared with the depth of the message in the threaded reply structure. The dialogue attributes that were defined based on timestamp order were not useful for discriminating between the constructs in either framework. That result was surprising because participants who came later to a discussion had the opportunity to read and reflect on all that had been written already before adding their own contribution. Previous work has shown that reading contributions from others is correlated with higher recorded performance (Wise et al., 2014) and with students having greater agency (Scardamalia & Bereiter, 2021). In the context of a threaded discussion assignment, it would thus seem that those who contribute later would have an advantage. Instead, the results in this chapter show that participants who added their message at the end of a chain of earlier contributions tended to write messages that demonstrated greater cognitive quality, regardless of when in time they were written, taking advantage of the “successively expanding history” (Suthers et al., 2010, p. 14) of messages and interactions as they create meaning.

Another aspect of the distinction between temporal order and thread depth that merits further investigation is the extent to which the apparent benefits of adding a
### Attribute Description

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positively correlated with cognitive quality</td>
<td></td>
</tr>
<tr>
<td><code>cm.DESWC</code></td>
<td>Number of words</td>
</tr>
<tr>
<td><code>cm.LDVOC</code></td>
<td>Lexical diversity, VOCD</td>
</tr>
<tr>
<td><code>message.depth</code></td>
<td>Message depth in discussion</td>
</tr>
<tr>
<td>Negatively correlated with cognitive quality</td>
<td></td>
</tr>
<tr>
<td><code>cm.LDTTRa</code></td>
<td>Lexical diversity, all words</td>
</tr>
<tr>
<td><code>cm.LDTTRc</code></td>
<td>Lexical diversity, content words</td>
</tr>
<tr>
<td><code>liwc.QMark</code></td>
<td>Number of question marks</td>
</tr>
<tr>
<td>Highest values for <strong>Affirmation</strong> messages (no indicators of cognitive presence)</td>
<td></td>
</tr>
<tr>
<td><code>liwc.affect</code></td>
<td>Number of affective process words</td>
</tr>
<tr>
<td><code>liwc.posemo</code></td>
<td>Number of +ve emotion words</td>
</tr>
<tr>
<td>Highest values for <strong>Triggering Event</strong> messages (<strong>Active</strong> and <strong>Constructive</strong> modes)</td>
<td></td>
</tr>
<tr>
<td><code>cm.DESWLltd</code></td>
<td>SD of word length in letters</td>
</tr>
<tr>
<td><code>message.replies.direct</code></td>
<td>Number of direct replies</td>
</tr>
<tr>
<td><code>message.replies.all</code></td>
<td>Total number of replies</td>
</tr>
</tbody>
</table>

Table 3.1: Dialogue attributes found to be most predictive for both the CoI phases of cognitive presence and the ICAP modes of cognitive engagement.
message at the end of a longer thread might be influenced by the role taken on by the participant. As the example interaction in Figure 2.1 illustrated, the student who started the thread by sharing their own presentation also assumed responsibility for managing the discussion. There was a strong tendency for that student to respond to every comment made by another participant – mainly to answer questions, but also to correct misunderstandings and to confirm or refute assertions from others. Proportionally, the participant who started the thread was responsible for around half of the conversational moves in every thread. The next chapter looks in more detail at the roles played by the participants during the discussion assignment and how the cognitive quality of their contributions varied across the four instructional scaffolding conditions.
Chapter 4

Integrating multiple quality measures

4.1 Introduction

This chapter and the following chapter are both concerned with the potential moderating influence of instructional interventions on measures of cognitive quality in online discussions. In the previous chapter, the two sets of frameworks constructs that define cognitive quality, the CoI phases of cognitive presence and the ICAP modes of cognitive engagement, were addressed separately. The main contribution of this chapter is a novel method for integrating multiple measures of quality, allowing us to use the insights from CoI and ICAP together to create a two-dimensional picture of cognitive quality.

An earlier study, focusing on “the depth and quality of learning rather than the quantity” (Schrire, 2006, p. 54), compared CoI with Bloom’s taxonomy (Bloom, 1956) and the SOLO taxonomy (Biggs & Collis, 1982) and found that there were broad correspondences between the constructs from the three frameworks. In particular, all three frameworks were in general agreement about the messages that were classified as exhibiting higher-order critical thinking skills, although they differed in their assessment of the remaining messages. The focus on “depth and quality of learning” in that study (Schrire, 2006) matched well with our focus on cognitive quality, and we decided to extend the comparison of frameworks to include ICAP, since no previous studies had compared CoI and ICAP in this way.

We adopted the same approach as Schrire (2006) to carry out an initial comparison of the indicators of higher-order and lower-order thinking in our data, except that in our case we used both CoI and ICAP. In common with that earlier work (Schrire, 2006), we grouped together the two highest phases of cognitive presence, Integration and Resolution, to form the higher-order group, while only Exploration was considered to
Chapter 4. Integrating multiple quality measures

### Table 4.1: Indicators of higher-order and lower-order thinking.

<table>
<thead>
<tr>
<th>CoI Group</th>
<th>ICAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution Integration</td>
<td>Higher-order</td>
</tr>
<tr>
<td>Exploration</td>
<td>Lower-order</td>
</tr>
<tr>
<td>Triggering Event Other</td>
<td>Other</td>
</tr>
</tbody>
</table>

The CoI phases of cognitive presence and the ICAP modes of cognitive engagement, grouped into indicators of higher-order and lower-order thinking, following Schrire (2006).

be lower-order (Table 4.1). The Integration and Resolution phases were also grouped together in earlier work (T. E. McKlin, 2004). The ICAP modes of cognitive engagement were not used in the Schrire (2006) study. For the purposes of the initial comparison, we chose to group Interactive and Constructive Reasoning together as higher-order, and to consider only Constructive Extending as lower-order. In previous work, the Constructive and Interactive modes were considered to be indicators of higher-order thinking behaviours (Wang et al., 2016b; Vellukunnel et al., 2017). We made use of the finer-grained distinctions in the extended ICAP taxonomy (Section 2.1.2) to separate the two sub-classes of Constructive mode, noting that the only difference between the Interactive and Constructive Reasoning modes is that a message can only be labelled as Interactive if it is a direct response to the substantive content of a previous message (Chi & Wylie, 2014; Yogevo et al., 2018).

We plotted the distribution of the higher-order and lower-order groups across the six course sessions represented in the data set of messages that was introduced in Chapter 2, where the CoI phases of cognitive presence and ICAP modes of cognitive engagement had been labelled by human annotators (Figure 4.1). As we discuss further in the next section, there was a change in the participation instructions between sessions 2 and 3 of the course, as detailed by Gašević, Adesope, et al. (2015). Participants in the later sessions received additional guidance for the discussion assignment, which was designed with the aim of improving the quality of their self-regulation through external facilitation and, by extension, increasing the cognitive...
4.1. Introduction

(a) Cognitive Presence

(b) Cognitive Engagement

Figure 4.1: Higher-order and lower-order thinking across six course sessions. Distribution of messages across the (a) higher and lower CoI phases of cognitive presence, and (b) higher and lower ICAP modes of cognitive engagement. The vertical line indicates the division between the Control and Treatment groups.

Looking first at cognitive presence, in Figure 4.1(a), there was a clear change in the distribution of messages between the Control and Treatment groups. In the Treatment group, there were consistently more messages displaying signs of the higher phases of cognitive presence, and the proportion of messages displaying the lower phases was reduced. This difference was previously confirmed as significant by Gašević, Adesope, et al. (2015) with the use of mixed-effects models. Thus, the difference between Control and Treatment groups, visible in Figure 4.1(a), could be seen to indicate that the additional guidance had fulfilled its aim of improving self-regulation and increasing discussion quality (Borge & Rosé, 2021). However, there was no such systematic effect on the distribution across the higher and lower modes of cognitive engagement, in Figure 4.1(b). In fact, the proportion of messages displaying higher and lower levels remained relatively stable across all six sessions. This preliminary result indicated that the associations between the constructs of the two frameworks
were not generally robust to the effects of the instructional interventions, and motivated us to examine the associations more closely.

The work reported in the rest of this chapter goes beyond a binary division into higher and lower levels of cognitive quality and examines the associations between each of the CoI phases of cognitive presence and the ICAP modes of cognitive engagement in more detail. Our second research question, as articulated in Section 1.1, asked about the robustness of those associations under different instructional scaffolding conditions. We used a combination of methods to address the question in two ways. First, we generated cross-tabulations and heat maps to measure the distribution of messages across each of the CoI phases of cognitive presence and the ICAP modes of cognitive engagement, plotting the construct labels against each other within the four instructional scaffolding conditions. The plots demonstrated visually that the associations between constructs were indeed moderated by the instructional interventions.

Second, we used a network analytic approach, where the co-occurrences of constructs from each framework could be examined using ENA. ENA was developed specifically for this type of analysis, where the number of concepts to be compared is much smaller than the number of interactions. The connections between concepts are the primary variable of interest: a single concept appearing on its own is of no value within ENA. In our data set, every message had exactly two labels: one from CoI and one from ICAP. We generated an initial ENA network using the pairs of labels and then compared that overall network against networks constructed from subsets of the data, defined by the instructional scaffolding conditions.

4.2 Peer-reviewed publication: A Network Analytic Approach to Integrating Multiple Quality Measures for Asynchronous Online Discussions

This section includes the verbatim copy of the following peer-reviewed publication, reprinted with permission:

Farrow, E., Moore, J., & Gašević, D. (2021a). A network analytic approach to integrating multiple quality measures for asynchronous online discussions. *LAK21: 11th*
Contributions: The ideas and analysis in the paper were developed and discussed between all authors of the work. The original idea, the experiments, and the writing were the work of the first author.
A network analytic approach to integrating multiple quality measures for asynchronous online discussions

Elaine Farrow
Johanna Moore
Elaine.Farrow@ed.ac.uk
J.Moore@ed.ac.uk
School of Informatics, University of Edinburgh
Edinburgh, UK

ABSTRACT
Asynchronous online discussions within a community of learners can improve learning outcomes through social knowledge construction, but the depth and quality of student contributions often varies widely. Approaches to assessing critical discourse typically use content analysis to identify indicators that correspond to framework constructs, that in turn serve as measures of depth and quality. Often only a single construct is addressed for performing content analysis in the literature, although recent work has used both social presence and cognitive presence constructs from the Community of Inquiry (CoI) framework. Nevertheless, there is no effective, commonly used, analytic approach to combining insights from multiple perspectives about quality and depth of online discussions. This paper addresses the gap by proposing the combined use of cognitive engagement (the ICAP framework) and cognitive presence (CoI), and by proposing a network analytic approach that quantifies the associations between the two frameworks and measures the moderation effects of two instructional interventions on those associations. The present study found that these associations were moderated by one intervention but not the other; and that messages labelled with the most common phase of cognitive presence could be usefully assigned to smaller meaningful subgroups by also considering the mode of cognitive engagement.

CSCS CONCEPTS
• Computing methodologies → Modeling methodologies; • Applied computing → Collaborative learning.

KEYWORDS
discussion forum, critical thinking, Community of Inquiry, cognitive presence, ICAP, Epistemic Network Analysis

1 INTRODUCTION
Text-based online discussion forums have been used in a wide variety of educational settings for many years [14, 35]. In the context of the recent sudden and substantial increase in online education brought about by a global pandemic, research into the learning that takes place as students interact in such a setting is both timely and important. The very act of composing and submitting messages to a forum of peers can directly lead to learning, particularly when it involves refining and negotiating meaning in cooperation with others [10]. Previous studies found that participation in discussion forums was positively correlated with learning gains in MOOCs, even in the cases where messages received no reply or only a superficial response [31, 32, 34]. This could be attributed to the known benefits of self-explanation [4], which does not require a conversational partner. However, both the ‘private world’ of reflection and the ‘shared world’ of discussion play a vital role in learning, according to the practical inquiry model [15, 17].

This study aims to develop a richer representation of the depth and quality of student participation in online discussion forums. In this, we follow Schrire in taking “a perspective emphasizing the depth and quality of learning rather than the quantity” [28, p. 54]. Specifically, this study brings together insights from two of the most widely-used and well-supported frameworks. We posit that the frameworks are related, yet different in important ways; and, therefore, that they provide complementary perspectives to the study of the depth and quality of participation. The Community of Inquiry framework (CoI) [16] was designed specifically to support online learning through computer-mediated discussion. Meanwhile, the ICAP framework [5] has been applied successfully in many different educational situations – classroom-based as well as online. Previous work has considered the link between certain dialogue attributes and measures of participation in these two frameworks [8], but did not look directly at how the constructs were related to one another. Analysing the correspondences between constructs from the two frameworks would allow us to develop a richer approach to the measurement of depth and quality in online discussions than using either one alone. In this work, our specific intent is to inform the development of a learning analytic approach to measuring the association between the frameworks and assessing whether that association is moderated by instructional scaffolds, given the importance of instructional guidance to support social knowledge construction in online discussions [13].
In the current study, we focused on the definitions of the phases of cognitive presence from CoI and the modes of cognitive engagement in ICAP. We contribute to the existing body of literature on assessing the depth and quality of student participation in online discussions in the following ways. 1) We theorize about the general associations between the constructs of two widely-used theoretical frameworks, grounding our arguments in empirical data. 2) We propose a novel network analytic approach to the analysis of associations between the two theoretical frameworks and demonstrate how it can be used in assessment of depth and quality of online discussions. 3) We examine the associations between the phases of cognitive presence and the modes of cognitive engagement and how they are moderated by institutional scaffolds aimed at promoting cognitive presence.

2 BACKGROUND

2.1 Analysis of social knowledge construction in online discussion

The analysis of social knowledge construction through collaborative asynchronous discussion can be investigated from multiple perspectives and at different levels of granularity: from coarse-grained patterns of interaction between participants, through message-based cognitive and social presence labels, to fine-grained discourse analysis looking at the individual dialogue acts and conversational moves within a message [28, 29]. The content of individual messages can be labelled, through content analysis [7], using various theoretical frameworks to identify expressions of critical thinking and conversational moves that support the establishment of a productive and supportive online community. Two of the most widely-used frameworks are the Community of Inquiry framework (CoI) [15, 16] and the ICAP framework [5], discussed in more detail below (Sections 2.2 and 2.3). Both frameworks define coding schemes that allow annotators to assign framework labels consistently to previously unseen content, for the purposes of content analysis. Recent work has identified cue phrases and dialogue features that are correlated with specific framework labels [8, 10], and relationships between the labels and topics extracted from the course content [27]. Although manual content analysis is slow and expensive, several recent studies have achieved promising results automating the labelling process using these frameworks [1, 9, 11, 20, 21, 23, 25].

Combining multiple perspectives on the analysis of online discussions can bring further benefits. One study [28] that used a multi-framework approach found correspondences between the patterns of interactions among participants, the phases of critical inquiry, and the conversational moves in the messages. In the subset of discussion threads that were characterized by direct interactions between students, more of the messages were labelled as belonging to higher phases of critical thinking, compared with threads where most messages were responses to the instructor’s original discussion prompt. Three different frameworks were used to assess the presence and extent of critical thinking. These were Bloom’s Taxonomy [3], the SOLO taxonomy [2], and CoI [15, 16], referred to in that work as the Practical Inquiry Model. Schrire acknowledged that triangulating findings across the three models is “conceptually problematic since each model is based on a different theory of cognitive activity” [28, p. 66]. Nevertheless, broad correspondences were confirmed, supporting social constructivist learning approaches. Whereas previous work [28, 31] looked in general terms at indicators of high and low levels of critical thinking, the present study looked into the detail of the individual constructs in each framework and examined how the associations between them were moderated by two instructional interventions.

2.2 The Community of Inquiry framework

The Community of Inquiry (CoI) framework identifies three ‘presences’ that are important for successful learning: social presence, cognitive presence, and teaching presence, through which the members of a community of learners connect with one another as ‘real people’ [16]: teaching presence, including the design of the course and assignments as well as ongoing facilitation; and cognitive presence, which relates to the students’ intellectual engagement with the course content. Cognitive presence is considered to be the element most basic to educational success – in fact, both social and teaching presence function primarily as a support for cognitive presence [16]. It aims to capture the process whereby discussion participants construct meaning through communication. Cognitive presence can be conceptualized as a measure of the depth and quality of student participation [5]. It has four phases:

- **Triggering Event:** a question, task, or problem that triggers the process of critical inquiry.
- **Exploration:** an exploration of the task or problem through the exchange of ideas, but lacking selectivity.
- **Integration:** an examination and integration of ideas through identifying connections and constructing meaning.
- **Resolution:** a resolution of the original problem, coupled with building consensus among participants.

Progression through the phases is expected to develop over time as the activities in each phase build on the previous one, although progress is seldom linear. The topic of the discussion is set by the initial Triggering Event. Messages in the Exploration phase tend to be wide-ranging, though always linked back to the original topic. The integration of ideas and the beginnings of a coherent line of reasoning signal that the discussion has entered the Integration phase. Many discussions will not reach the Resolution phase without intervention from an instructor, since students often feel more comfortable in the earlier phases [17]. Resolution can involve consensus building within the community and the generation of possible solutions to the original problem, sometimes in the form of thought experiments. Clarification questions and new ideas can appear at any point in the discussion, perhaps beginning a new cycle. Messages containing indicators from multiple phases are coded with the highest phase (“coding up”) [33].

2.3 The ICAP framework

Cognitive engagement is the central concept in the ICAP framework [5]. The framework defines four modes of cognitive engagement, based on observable student behaviours: Interactive, Constructive, Active, and Passive.

- **Interactive:** an interaction with a partner while both are engaged constructively.
- **Constructive:** the generation of novel content; for example, through reasoning or summarization.
• Active: an activity that demands attention, such as referencing previously given information.
• Passive: an activity where attention can wander; for example, reading without responding.

Higher modes are theorised to correlate with greater learning gains. Each mode represents a qualitative shift in knowledge growth. The original framework was adapted and expanded in several recent studies \cite{8, 31, 36}. The label definitions for Constructive and Active were subdivided to introduce Constructive Reasoning and Constructive Extending, and Active Targeted and Active General, respectively. Messages consisting primarily of agreement or thanks were also given their own label, Affirmation \cite{8}. In the context of online discussions, the Interactive label was reserved for messages that contained a direct response to the content of an earlier contribution. Responses to external information sources, such as textbooks or video presentations were instead labelled as Constructive.

The extended ICAP schema was used to label messages in MOOC discussion forums \cite{31, 32}, and on MOOC videos \cite{36}, although in the latter case the activity was carried out individually with no scope for interaction between participants. Results from these studies indicated that interaction was rare and the majority of messages did not receive contentful responses. The students who contributed the most frequently tended to generate questions rather than building on what others had written. An intervention to encourage students to generate more content \cite{36}, simply increased the proportion of shallow and simple (Active mode) messages and did not improve learning gains.

3 RESEARCH QUESTIONS

Section 2.1 presented the benefits of bringing together a variety of perspectives in order to achieve deeper understanding of how social knowledge construction takes place through asynchronous discussion. Previous studies achieved new insights by utilising both coarser-grained and fine-grained analyses \cite{28} and by looking at the relationship between different presences within CoI \cite{26}. There has been less research bringing together insights from multiple theoretical frameworks. In particular, the existing research has little to say about how the two most widely-used frameworks for assessing the depth and quality of student participation in online discussion forums are related. The first research question in the present study was therefore:

**RQ1. Are the general associations between the individual CoI phases of cognitive presence and the ICAP modes of cognitive engagement?**

The primary role of teaching presence is to support the development of cognitive presence \cite{15, 16}, so it is important to evaluate the effect of different instructional scaffolds on the development of concepts from the two frameworks and the associations between them. It is easy to imagine how changes in knowledge construction prompted by an intervention could improve one metric, while leaving another unaffected or even reduced – for example, by increasing the quantity of simple questions asked \cite{30}. In this study, we specifically focused on two types of intervention to examine whether and how they moderated the associations between CoI and ICAP. One intervention involved the assignment of specific roles within the discussion, which is a common scaffolding approach in computer-supported collaborative learning \cite{13}. The other one was based on increasing the scaffolding support to encourage deeper and higher quality engagement through providing quality standards that quality contributions to online discussion should have. Prior work showed that both of these interventions were effective in improving measures based on the CoI phases of cognitive presence \cite{18}, and that they had a moderating effect on the associations between cognitive presence and social presence \cite{26}, and between social presence and learning outcomes \cite{22}. However, no analysis had looked at the effect on the ICAP modes of cognitive engagement, or any changes that might be seen in the patterns of connection between those and the CoI phases of cognitive presence. Thus, our second research question was:

**RQ2. Are, and if so, how are the associations between the CoI phases of cognitive presence and the ICAP modes of cognitive engagement moderated by the instructional scaffolds aimed at promoting cognitive presence?**

4 METHOD

4.1 Description of the data

The course discussion forum messages used in this study were collected during six sessions of the same online Masters level course at a Canadian university between 2008 and 2011. Each message thread corresponded to a discussion that was led by a single student, who took on the role of Research Expert for the duration of the thread. The opening post in each thread introduced the new topic and included a link to a video presentation the student had prepared based on their own choice of research paper. The other students in the thread took the Practising Researcher role, giving feedback on the presentation and asking questions about the content. The course instructors rarely got involved with the discussions – only 6 posts in total across the whole data set of 1,747 messages were written by an instructor. Forum participation was graded and accounted for 10% of the overall course mark.

4.2 Framework labels from CoI and ICAP

The messages were labelled with both the CoI phases of cognitive presence and the ICAP modes of cognitive engagement. Exactly one label from each framework was used for each message. If a message matched the description for more than one label in either framework, the label that corresponded to greater depth and quality of participation was chosen. The additional labels Other and Off-task were used where there were no indications of cognitive presence or cognitive engagement, respectively. They are not generally considered to be part of the frameworks.

The labels for the CoI phases of cognitive presence were assigned by two expert coders who achieved high levels of agreement (98.1% agreement, Cohen’s $κ = 0.974$). The label distribution is shown in Table 1. In a similar way, the cognitive engagement labels of ICAP were assigned by two postgraduate students working independently, achieving ‘substantial’ inter-annotator agreement (Cohen’s $κ = 0.623$) \cite{24}. An expanded set of ICAP labels was used, following earlier work \cite{8, 31, 32, 36} (see Section 2.3). The distribution of these labels across the data set is presented in Table 2.
Table 1: Breakdown of messages by Col phases of cognitive presence in the whole data set. The Other label was used where a message did not display any cognitive presence.

<table>
<thead>
<tr>
<th>Cognitive presence phase</th>
<th>Short label</th>
<th>Example behaviour</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triggering Event</td>
<td>TRIGGERING</td>
<td>Asking a question or posing a problem</td>
<td>508</td>
<td>57.06%</td>
</tr>
<tr>
<td>Exploration</td>
<td>EXPLORATION</td>
<td>Exchanging ideas</td>
<td>684</td>
<td>39.15%</td>
</tr>
<tr>
<td>Integration</td>
<td>INTEGRATION</td>
<td>Integrating ideas and constructing meaning</td>
<td>568</td>
<td>29.05%</td>
</tr>
<tr>
<td>Resolution</td>
<td>RESOLUTION</td>
<td>Reaching consensus or suggesting a new hypothesis</td>
<td>187</td>
<td>6.12%</td>
</tr>
<tr>
<td>Other</td>
<td>OTHER</td>
<td>Commenting with no signs of cognitive presence</td>
<td>140</td>
<td>8.01%</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
<td>1,747</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Table 2: Distribution of cognitive engagement modes in the whole data set. The Off-task label was used for messages displaying no cognitive engagement. The Passive label was not used in this data set.

<table>
<thead>
<tr>
<th>Cognitive engagement mode</th>
<th>Short label</th>
<th>Example behaviour</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive</td>
<td>I</td>
<td>As for C1, but in direct response to an earlier message</td>
<td>579</td>
<td>33.14%</td>
</tr>
<tr>
<td>Constructive Reasoning</td>
<td>C1</td>
<td>Displaying explanation or reasoning about the current topic</td>
<td>513</td>
<td>32.92%</td>
</tr>
<tr>
<td>Constructive Extending</td>
<td>C2</td>
<td>Introducing new content to the discussion</td>
<td>675</td>
<td>43.81%</td>
</tr>
<tr>
<td>Active Targeted</td>
<td>A1</td>
<td>Referencing specific previous content</td>
<td>75</td>
<td>4.70%</td>
</tr>
<tr>
<td>Active General</td>
<td>A2</td>
<td>Showing other signs of being engaged with course content</td>
<td>287</td>
<td>16.57%</td>
</tr>
<tr>
<td>Passive</td>
<td>P</td>
<td>Reading messages without responding</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Affirmation</td>
<td>F</td>
<td>Affirming what was said in an earlier message</td>
<td>75</td>
<td>4.41%</td>
</tr>
<tr>
<td>Off-task</td>
<td>O</td>
<td>Commenting with no relation to the current topic or the course</td>
<td>11</td>
<td>0.63%</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
<td>1,747</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

4.3 Scaffolding interventions

Two distinct scaffolding interventions took place during this course. One was the within-subjects role assignment described in Section 4.1, where each student took a turn at being the Research Expert and presented their work to a group of Practicing Researchers. The second intervention was a change in the participation instructions between the second and third sessions of the course, with additional guidance given about what would constitute a high quality contribution to the discussion (Figure 1). The first two sessions of the course thus constitute the Control group, while the remaining four sessions, where the additional guidance was given, are the Treatment group. The number of participants and messages exchanged in each instructional condition is shown in Table 3.

Table 3: Messages sent and unique participants in each instructional condition. Students were expected to take on both the Research Expert and Practicing Researcher roles. The instructor and one student who repeated the course participated in both Control and Treatment groups.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Treatment</th>
<th>Total</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practising Researcher</td>
<td>414</td>
<td>464</td>
<td>968</td>
<td>81</td>
</tr>
<tr>
<td>Research Expert</td>
<td>411</td>
<td>436</td>
<td>847</td>
<td>82</td>
</tr>
<tr>
<td>Total</td>
<td>825</td>
<td>902</td>
<td>1,747</td>
<td>85</td>
</tr>
</tbody>
</table>

4.4 Data analysis methods

4.4.1 Cross-tabulation. As a first step towards answering each of our research questions, we used a cross-tabulation approach to examine the distributions of the framework labels across the data set, in a similar manner to prior work [28]. Plots such as bar charts and heat maps are useful for visualising the results of cross-tabulation, in order to obtain a preliminary understanding of the data set as a whole, relevant to RQ1. They also allow for comparing different subsets, relevant to RQ2.

4.4.2 Epistemic Network Analysis. The second stage of our analysis used Epistemic Network Analysis (ENA) [29]. ENA is a network analytic approach that is ideal for investigating the relationships between relatively small sets of concepts across a densely connected network of interactions. It has been used successfully in many previous studies of cognitive presence and social presence [12, 19, 26, 27]. Different sub groups of analysis units (for example, students or messages) can be compared both visually and statistically, providing both qualitative and quantitative insights.

ENA is based around measuring and visualising the network of interactions. It has been used successfully in many previous studies of cognitive presence and social presence [12, 19, 26, 27]. Different sub groups of analysis units (for example, students or messages) can be compared both visually and statistically, providing both qualitative and quantitative insights.
is projected down onto a lower-dimensional projection space using singular value decomposition (SVD). Network edges are weighted by the frequency of connections between pairs of concepts. This method ensures that data points with similar patterns of connections will generally appear close together in the projection space, while those that differ more will appear further apart. Additionally, the positioning of the concepts as nodes within the projection space can indicate which concepts tend to be linked to other concepts in similar ways, and thus can make the space itself interpretable [29].

When the network is projected down to the two most informative dimensions (that is, into a flat 2D representation), it is possible to visualise the network of connections for a single data point – for example, one student. Each individual network can also be summarised by a single point in the projection space that represents the centroid, or weighted mean, of that network. This allows comparisons on a larger scale, for example across a cohort of students.

The fundamental building blocks of the ENA approach are the unit of analysis and the conversation (or stanza). The unit of analysis defines how conversations are grouped together to produce the data points for the network, while the choice of conversation determines which connections are included in a particular analysis. All concepts that appear anywhere within the same conversation are considered to be connected to one another. Typically, no weighting is used at this level – concepts are either connected or not. A conversation can be as short as an individual message, or much longer, for example, an entire discussion thread. The unit of analysis combines conversations to form the weighted network. For example, the conversations on each day might be grouped together for analysis, in order to track the development of conceptual links over time, with the unit of analysis being a day rather than a student.

We used the same network configuration to answer both of our research questions. We set the conversation parameter to be the message, allowing us to examine the association between framework concepts at the lowest granularity. Since we intended to look at the moderating effect of the instructional scaffolds in RQ2, the unit of analysis was a compound one: student within instructional condition. Students typically belonged to only one of the Control and Treatment groups but took on both the Research Expert and Practising Researcher roles. By aggregating together the messages sent by a student within a single instructional condition, each student was represented in the network data twice, once for each role. We excluded from the ENA analysis the 6 messages that were sent by a student within a single instructional condition, each student tended to repeat the course and thus did not act as the Research Expert.

As RQ1 aimed to better understand the general associations between the individual CoI phases of cognitive presence and the ICAP modes of cognitive engagement, we explored the mean network formed by all students taken together. Our first research question asked about the general associations between the framework constructs.

5.1 Results for RQ1: general associations between the framework constructs

Our first research question asked about the general associations between individual labels in each of the frameworks; that is, between the CoI phases of cognitive presence and the ICAP modes of cognitive engagement. The distribution of labels corresponding to each of the CoI phases of cognitive presence was plotted against the labels from the extended ICAP taxonomy in Figure 2. No exclusive correspondences were found between any pairs of labels. However, a trend was visible in the chart whereby the labels indicating greater depth and quality of participation in the CoI framework appeared more frequently in conjunction with the higher ICAP indicators. For example, the number of messages with the Integration label increased almost linearly across the three highest ICAP modes, while Triggering Event messages were approximately equally likely to belong to Active Targeted and Constructive Extending.

Figure 3 shows the average ENA network for all students, using data from all four instructional conditions. The framework labels are shown using their short labels (Tables 1 and 2) to reduce visual clutter. The X axis accounts for 30.8% of the variance in the data and the Y axis accounts for 11.5%. The X axis primarily distinguishes between the interactive and non-interactive ICAP modes of cognitive engagement, while the Y axis distinguishes between the early phases (Triggering Event and Exploration) and the later phases (Integration and Resolution) of cognitive presence.

Table 4 presents the strength of each network connection from Figure 3. There were no links among pairs of codes within the same framework because each conversation (a message) had exactly one label from each framework. The strongest connections were found between the higher-order indicators across the two frameworks. The strongest of all was the link between Interactive mode and Integration phase at 0.31, followed by Interactive mode and Exploration phase at 0.22. This result is in line with the expectation that messages in both of these CoI phases of cognitive presence tend to build on the content of previous discussion contributions, as required by the definition of ICAP Interactive mode.

5.2 Results for RQ2: the potential moderating role of instructional scaffolds

5.2.1 Cross-tabulation. The results in Section 5.1 gave a high-level overview of the whole data set and the general associations between the framework concepts. Our second research question asked about

---

1There were a small number of exceptions. One student repeated the course and thus participated in both the Control and Treatment groups. Four students never took on the Practising Researcher role and two did not act as the Research Expert.
Figure 2: Message counts for each of the CoI phases of cognitive presence, broken down by the extended ICAP labels.

Table 4: ENA network weights for the network in Figure 3. Values greater than 0.10 are shown in bold.

<table>
<thead>
<tr>
<th></th>
<th>Other</th>
<th>Triggering Event</th>
<th>Exploration</th>
<th>Integration</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.00</td>
<td>0.00</td>
<td>0.22</td>
<td>0.31</td>
<td>0.09</td>
</tr>
<tr>
<td>C1</td>
<td>0.02</td>
<td>0.17</td>
<td>0.20</td>
<td>0.09</td>
<td>0.00</td>
</tr>
<tr>
<td>C2</td>
<td>0.05</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>A1</td>
<td>0.04</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Figure 3: ENA network constructed by summing the label co-occurrences within the messages sent by each individual student in each condition and plotting the normalised distributions. The node positions indicate the locations of each framework construct in the projected space. The thickness of the lines joining the nodes indicates the average strength of the connection across all data points.

5.2.2 ENA network for the whole data set. The ENA projection network in Figure 5 shows a collection of points corresponding to individual students, located based on the position of the centroid of that student’s network in the projection space, and coloured according to the relevant instructional condition. The squares indicate the group means, and the dashed lines around them represent the 95% confidence intervals. The Practising Researcher and Research Expert groups were separated along the X axis, corresponding to the divide between interactive and non-interactive ICAP modes, while the Control and Treatment groups were separated along the Y axis, corresponding to the split between the earlier and later CoI phases of cognitive presence. The separation between user roles

the potential moderating effects of two instructional scaffolds. To drill down into more detail we plotted heat maps for each of the four instructional conditions (Figure 4). The same general trend was seen across all four conditions as for the data set overall: the messages that received a higher indicator of depth and quality of participation on one framework were more likely to receive a higher indicator in the other framework as well, but there were no strong links between constructs at the lower levels.

One of the most striking differences between the conditions was the virtual absence of messages with the Constructive Reasoning label when the student was in the role of Research Expert, while such messages were commonly seen for students in the Practising Researcher role. The label definitions provide an explanation for this disparity. A message could only be labelled as Interactive if it was a response to the content of an earlier message, in addition to displaying explanation or reasoning. The first message in each thread in this data set lacked substantive content: the student simply announced the new topic and provided the link to their presentation. Thus, a message posted by a Practising Researcher in response to the opening message of the thread, and displaying explanation or reasoning, would be labelled as Constructive Reasoning rather than Interactive. In contrast, replies from the Research Expert frequently built on the content of earlier posts (Interactive) and only rarely got the Constructive Reasoning label.
was visually almost complete, with only a couple of data points from each role condition straying over the mid line. The distinction between the Control and Treatment groups appeared to be somewhat less clean. However, a series of Mann-Whitney tests showed many statistically significantly differences with large effect sizes (Table 5). In order to understand the impact of the instructional scaffolds more fully, we addressed them separately.

5.2.3 The effect of external facilitation. Separate ENA networks for the Control and Treatment groups were plotted (Figure 6) to visualise how the associations between the framework concepts varied with the external facilitation conditions. The network configuration parameters (unit of analysis and conversation) were the same as for the network built from the full data set (Figure 3), but the differences in the data resulted in the construction of different projection spaces. The amount of variance explained by the two major axes in each case was similar to the network in Figure 3 that used all the data (34.9% and 34.0% respectively for the X axis, and 12.2% and 11.8% for the Y axis). The absolute positions of the network nodes cannot be meaningfully compared because the networks each use a different projection space, but the relative positions of nodes with respect to one another within each network can be used to discover how the patterns of association between constructs differ between conditions. Specifically, the change in relative positioning of the nodes between the Control and Treatment networks indicates a shift in the associations between framework concepts due to the external facilitation intervention.

The general orientation of the networks in Figure 6 was similar to those in Figure 3, with the Constructive Reasoning and Interactive modes at opposite ends of the X axis, and the Y axis distinguishing between the earlier and later phases of cognitive presence. In the Control condition, the Constructive Reasoning mode was located near the Triggering Event phase; while in the Treatment condition it was closer to the Resolution phase. This suggests the existence of an association in the Treatment condition between using explanation and reasoning, and reaching the later phases of cognitive presence; while in the Control condition, despite sending such messages, students were more likely to remain in the earlier phases of cognitive presence. The Constructive Reasoning mode was used almost exclusively by students in the Practising Researcher group (Section 5.2.1), suggesting that the external facilitation intervention was successful in changing the focus of their contributions from exploring a single idea from the presentation (albeit with explanation and reasoning), to reasoning about how the presented material might relate to other content from the course.

Additionally, the Constructive Extending mode changed position from the right side of the network (more interactive) in the Control condition to the left side (less interactive) in the Treatment condition, while remaining nearer the early phases of cognitive presence than the later ones. One interpretation of this result is that the
Table 5: Results from a series of Mann-Whitney tests comparing conditions for statistical differences at the $\alpha = 0.05$ level along both X (SVD1) and Y (SVD2) axes, with significant results highlighted in bold. Effect sizes greater than 0.5 are considered large, while effect sizes between 0.3 and 0.5 are considered medium [6].

<table>
<thead>
<tr>
<th>Axis</th>
<th>Condition</th>
<th>Comparator</th>
<th>U</th>
<th>$p$</th>
<th>$\tilde{z}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD1</td>
<td>Control-Practising Researcher</td>
<td>Control-Research Expert</td>
<td>2.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Treatment-Practising Researcher</td>
<td>Control-Research Expert</td>
<td>4.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Treatment-Practising Researcher</td>
<td>Treatment-Research Expert</td>
<td>6.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Control-Research Expert</td>
<td>Treatment-Research Expert</td>
<td>3.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Treatment-Research Expert</td>
<td>Treatment-Research Expert</td>
<td>3.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Treatment-Research Expert</td>
<td>Control-Research Expert</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Control-Research Expert</td>
<td>Control-Practising Researcher</td>
<td>720.00</td>
<td>0.22</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Control-Research Expert</td>
<td>Treatment-Practising Researcher</td>
<td>353.00</td>
<td>0.00</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>Control-Research Expert</td>
<td>Control-Research Expert</td>
<td>386.00</td>
<td>0.00</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Treatment-Research Expert</td>
<td>Treatment-Research Expert</td>
<td>740.00</td>
<td>0.04</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Control-Practising Researcher</td>
<td>Control-Research Expert</td>
<td>391.00</td>
<td>0.00</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Control-Research Expert</td>
<td>Control-Research Expert</td>
<td>493.00</td>
<td>0.00</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Control-Research Expert</td>
<td>Control-Practising Researcher</td>
<td>657.50</td>
<td>0.63</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Figure 5: The ENA network constructed using a compound unit of analysis – student within instructional condition – with message as the conversation. It uses the same projection space as Figure 3 and shows the centroid of each student’s network as a point. To reduce visual clutter, the Practising Researcher group is labelled ‘Practitioner’ and the Research Expert group is shown as ‘Expert’.

students in the Control group, in both user roles, who tended to introduce new information to the discussion (Constructive Extending) were also likely to build on the contributions of others (Interactive), while in the Treatment group, the external facilitation intervention discouraged students from introducing new material without explaining how it linked to the earlier discussion – the behaviour captured by the combination of the Constructive Extending mode and the earlier phases of cognitive presence – and thus increased the differences between students who sent a lot of Interactive messages and those whose messages generally indicated lower levels of critical thinking.

5.2.4 The effect of assigning user roles. The Practising Researcher and Research Expert groups were plotted as separate ENA networks (Figure 7) to visualise how the associations between the framework concepts were moderated by the assignment of user roles. The configuration parameters were again the same as for the network constructed from the full data set (Figure 3), but the differences within the data resulted in different latent space projections – in this case leading to very different layouts compared to the earlier networks.

The network generated from the messages sent by students in the Practising Researcher role bore little resemblance to any of the earlier networks. The X axis accounted for 25.6% of the variance in the data and the Y axis for 18.1%. Unlike the earlier networks, the X axis primarily distinguished between the phases of cognitive presence, with the Triggering Event phase on the far right, the Exploration phase near the middle, and the Integration phase at the far left. The Resolution phase was used rarely in this subset of the data and appeared in the mid-left. The Y axis accounted for the distinction between the lower (Active General and Active Targeted) and higher (Constructive Extending, Constructive Reasoning, and Interactive) ICAP modes of cognitive engagement.

The Research Expert network shared some commonality with the Practising Researcher network. The phases of cognitive presence were again distributed along the X axis, accounting for 33.9% of the variance, although the positions of the Triggering Event and Exploration phases on the right of the network were reversed. The Y axis accounted for 17.0% of the variance. The Integration and Exploration phases were nearest the top, while the Resolution phase was at the bottom. Neither axis had a clear interpretation in terms of the ICAP modes of cognitive engagement.

We compared the relative positions of the individual constructs in the networks based on user roles in Figure 7 and the network constructed from the full data set in Figure 3. We found no clear indications of any changes to the associations between framework concepts moderated by the user role assignment intervention. In both conditions, the Exploration phase of cognitive presence was positioned near the ICAP Constructive Extending mode. In the Practising Researcher network, the Constructive Reasoning mode was also in the neighbourhood, while in the Research Expert network, the Interactive mode was nearby. This reflected the major difference in label distribution between these two conditions (Section 5.2.1) due to the requirement for an Interactive message to be a response.
to the substantive content of another message. Although the Constructive Reasoning mode in the Practising Researcher group was strongly linked to both the Exploration and Integration phases of cognitive presence, it was plotted in the same area of the network as Exploration and distant from Integration, indicating a greater similarity to Exploration. In a similar way, the Interactive mode in the Research Expert group was located near the Exploration phase, despite being strongly linked to both Exploration and Integration. This indicates that students whose messages tended to contain explanation or reasoning were nevertheless more likely to remain in the Exploration phase than to progress to Integration. Finally, the Constructive Extending mode was positioned near the centre in both networks, indicating that it represented the ‘conceptual average’ along both axes.

6 DISCUSSION

6.1 General associations between constructs

Our analysis of the general associations between the individual CoP phases of cognitive presence and the ICAP modes of cognitive engagement in RQ1 revealed a trend whereby the indicators of greater depth and quality of participation tended to be related while the same was not true of the indicators of lower depth and quality. This finding was evidenced both by cross-tabulation (Figure 2) and
by the construction of an ENA network from the framework labels assigned to each individual message (Figure 3). This result agrees with the findings of prior work that used CoI along with Bloom’s taxonomy and the SOLO taxonomy and found that the lower levels of the frameworks were “of a different nature” and thus could not be sensibly compared [28, p. 63]. Specifically, the lowest phase of cognitive presence, *Triggering Event*, indicates a message that can launch the discussion in a new direction, rather than moving towards a resolution of the current problem. In contrast, the lower modes of ICAP differ from the higher modes primarily because the contribution lacks novelty. Meanwhile, the two highest phases of cognitive presence involve bringing ideas together and constructing meaning (*Integration*) and reaching consensus (*Resolution*). The link between these and the higher modes of ICAP – *Constructive* activity to generate novel content and *Interactive* messages building on earlier contributions – is clearer.

That leaves the most common phase of cognitive presence, the *Exploration* phase, which can encompass anything from a simple paraphrase of previously shared content, through the introduction of new content, to an explanation that builds on the contribution of a partner but does not integrate multiple viewpoints. Incorporating the ICAP modes into our analysis allowed us to distinguish between these cases and thus develop a more nuanced appreciation of a student’s contribution to the discussion. Future work also incorporating social presence could develop this further to look at links between the ‘social climate’ of the discussion [26] and the ways students build on each other’s contributions.

### 6.2 Moderating role of instructional scaffolds

The second research question in this study, RQ2, aimed to explore the effects of two instructional interventions, described in Section 4.3. In each of the four conditions, the most common pairing of labels from the two frameworks was different (Figure 4).

- In the *Control-Practising Researcher* condition, where neither of the interventions was in effect, the most common label co-occurrence was the *Triggering Event* phase coupled with *Active Targeted* mode – a question relating back to information that was shared earlier in the discussion.
- In the *Control-Research Expert* condition, the most common pairing was the *Exploration* phase and *Interactive* mode – the exchange of ideas with a conversational partner – suggesting that students in the *Research Expert* role were active in driving the discussion forward.
- The most common pairing in the *Treatment-Practising Researcher* condition was the *Integration* phase with *Constructive Extending* mode, indicating the construction of meaning through explanation or reasoning. The influence of the revised assignment instructions is clear in shifting the behaviour of these students from simple clarification questions to more substantive contributions.
- Finally, in the *Treatment-Research Expert* condition, the most common co-occurrence was the *Integration* phase with *Interactive* mode, indicating the construction of meaning through explanation or reasoning in direct response to a previous message. Here, the effects of both interventions were combined, and students in the *Research Expert* role benefited from the change in behaviour of their conversational partners. This allowed the discussion to move from the *Exploration* to the *Integration* phase more frequently.

While the present study confirmed the significant shift from earlier to later phases of cognitive presence between the *Control* and *Treatment* groups in the between-subjects external facilitation intervention, noted in prior work [18, 26], there was no corresponding change in the distribution of lower to higher ICAP modes of cognitive engagement. ENA networks were used to examine and explain how the associations between the phases of cognitive presence and the ICAP modes were affected by the intervention, leading to the disparity in metrics (Figure 6).

In contrast, the role assignment intervention did not appear to have a clear moderation effect on the associations between the constructs of the two frameworks. The measures of depth and quality of participation on both frameworks were higher for the *Research Expert* group than for the *Practising Researcher* group. Further research might nevertheless reveal a moderating effect, as has been seen in work incorporating social presence [22, 26]. Future work should also consider how the association between the measures varies over time [31] and whether there is an ordering effect based on the students’ experience of the different roles.

### 6.3 Limitations

The main limitation of this study is that it used data from only a single course, although this was collected from several offerings across an extended period of time. The relationships between the theoretical frameworks that were discovered in this study might not hold for data from a different setting; for example, the fact that participation in the forum contributed to the course grade will undoubtedly have influenced the approach taken by students.

### 7 CONCLUSION

The primary research contribution offered by this study is a novel approach for exploring the relationship between different indicators of the depth and quality of participation in online discussion forums. By combining insights from a combination of cross-tabulation and network analytic techniques, we uncovered connections between indicators of CoI cognitive presence and ICAP cognitive engagement and saw how student behaviour was affected by two different types of instructional intervention. The method can be applied to other situations where multiple indicators interact in potentially complex ways.

A second contribution of this study is the evaluation of how the combination of external facilitation and role assignments affected the relationship between two different measures of the depth and quality of student participation, where positive changes in one measure were not always reflected in another measure. The use of two complementary perspectives in future studies is likely to become increasingly feasible, thanks to the development of automated classifiers for both CoI [9, 11, 21, 23, 25] and ICAP [1, 20, 36]. Combining the analytic approach presented here with real-time automated labelling could allow researchers and practitioners to benefit from rich analytic insights and to evaluate interventions while a course is still in progress.
ACKNOWLEDGMENTS

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


4.3 Summary of contributions

In this chapter, we presented a novel approach for combining insights from two distinct educational frameworks. We used that approach to examine the potential moderation effect of two different instructional interventions on the associations between the CoI phases of cognitive presence and the ICAP modes of cognitive engagement, in order to gain a two-dimensional understanding of cognitive quality. We answered our second research question from Section 1.1 by using a network analytic approach to examine how the associations between the CoI phases of cognitive presence and the ICAP modes of cognitive engagement were moderated by different instructional scaffolding conditions.

The way we used ENA in this chapter was novel in two respects. First, the labels came from two distinct, non-overlapping sets, and there were no connections between labels in the same set. Every message had exactly one label from each framework, constraining the number of possible connections.

Second, and more significantly, we constructed multiple networks using different subsets of the data. A more common approach is to start by constructing an ENA network using the whole data set (as we did in Figure 3 of Farrow et al. (2021a), reproduced on p. 93) and then to subtract the weights of the edges in one subgroup from those in another. The resulting comparison plot of the subtracted network indicates which of the connections are more significant in each subgroup. The differences between the subgroups can be indicated visually by colouring the remaining edges according to subgroup membership. For completeness, the subtracted networks for the two comparisons examined in this chapter are shown in Figure 4.2. In each case, the projection space and the explained variance were the same as for the network constructed from the full data set (Figure 3 in Farrow et al. (2021a), on p. 93).

The thick red and blue edges in Figure 4.2(a) relate to the way the Interactive mode was used almost exclusively by the Research Expert group, while the Constructive Reasoning mode played a similar role for the Practising Researcher group. The stark left-right division directly corresponds to the distribution of blue and red points in Figure 5 of Farrow et al. (2021a), reproduced on p. 95, where the strong distinction was noted and discussed.

The green and purple edges in Figure 4.2(b) correspond to the expected shift from earlier to later CoI phases of cognitive presence, as a result of the external facilitation intervention (Gašević, Adesope, et al., 2015; Rolim, Ferreira, et al., 2019).
The connections that remain after subtraction are generally weaker than those in Figure 4.2(a). The role assignment intervention led to a large difference between the messages sent by participants in the Practising Researcher and Research Expert roles. The change in external facilitation between Control and Treatment groups had a smaller, but still noticeable, effect.

The subtracted networks in Figure 4.2 were useful to visualise how the strength of the associations between framework constructs varied across instructional conditions, but they could not easily show changes in the patterns of association, since they all used the same projection space as the full network. It was for that reason that we generated entirely new networks for each subset of the data (Figures 6 and 7 in Farrow et al. (2021a), reproduced on p. 96). Their coordinate spaces were different, so these networks were not directly comparable with one another. Nevertheless, we identified broad similarities in structure between the networks generated for the Control and Treatment groups and the overall network defined by the full data set. In contrast, the two networks defined using only the messages sent while participants were in either the Research Expert role or the Practising Researcher role bore little resemblance to one another or to the network constructed from the full data set.

The next chapter looks more closely at the role assignment intervention, paying particular attention to possible effects of the order in which participants experienced each role.
4.3. Summary of contributions

(a) Comparison plot showing the Practising Researcher group, in blue, and the Research Expert group, in red.

(b) Comparison plot showing the Control group, in green, and the Treatment group, in purple.

Figure 4.2: Subtracted networks for each intervention, showing differences between the mean networks for (a) role assignment, and (b) external facilitation. The networks were constructed using the connections between the CoI phases of cognitive presence and the ICAP modes of cognitive engagement within individual messages in our data set.
Chapter 5

Effects of temporal ordering

5.1 Introduction

The previous chapter investigated the influence of two instructional interventions on patterns of association between the indicators of cognitive quality in online discussions. The present chapter looks in more detail at the effects on cognitive quality of assigning predefined roles to the discussion participants. In particular, it focuses on possible differential effects that could be attributable to the effect of experiencing the roles in a particular order, and thus answers the third research question from Section 1.1.

Many experimental settings require participants to be divided into groups, and for those groups to experience different treatments or interventions. In the case of assigned roles, rotating the participants through the roles is common (Schellens et al., 2007; Wise et al., 2012) and is generally considered the fairest approach (Strijbos & Weinberger, 2010). Nevertheless, care must be taken to ensure that none of the participants is disadvantaged by the chosen order. This is of particular importance when the intervention is conducted in the context of a credit-bearing course. For example, one study (De Wever et al., 2010) considered a group task that comprised four separate discussions: half of the groups were given assigned roles for the first two discussions (only), while the other half were only assigned roles during the final two discussions. The groups that were given defined roles at the outset performed better overall, despite the roles being optional during their later discussions, indicating that early experience of the roles provided an ongoing benefit.

Many different roles have been defined in earlier studies, both emergent (Strijbos & De Laat, 2010; Strijbos & Weinberger, 2010; Dowell & Poquet, 2021; M. Ferreira et al., 2021; M. Ferreira et al., 2022) and assigned (Schellens et al., 2007; De Wever et al.,
Table 5.1: Single duty roles used in previous studies. Topic Leader is a more specialised type of Starter, requiring greater knowledge of the topic. Synthesizer and Wrapper are both sub-types of Summariser.

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starter</td>
<td>Introduce the topic and keep the discussion going by adding new points</td>
</tr>
<tr>
<td>Topic Leader</td>
<td>Introduce the topic and provide knowledgeable input throughout the discussion</td>
</tr>
<tr>
<td>Moderator</td>
<td>Monitor and guide the discussion, asking and answering questions and providing feedback</td>
</tr>
<tr>
<td>Source Searcher</td>
<td>Identify and contribute information from outside sources to support or challenge contributions from others</td>
</tr>
<tr>
<td>Theoretician</td>
<td>Make claims, supported by evidence; provide missing information; ensure all relevant theories are considered</td>
</tr>
<tr>
<td>Summariser</td>
<td>Synthesise and summarise contributions from others, incorporating and highlighting different viewpoints</td>
</tr>
<tr>
<td>Synthesizer</td>
<td>Produce an interim summary, highlighting agreements and disagreements and suggesting next steps</td>
</tr>
<tr>
<td>Wrapper</td>
<td>Produce a final response on behalf of the group, integrating different perspectives</td>
</tr>
</tbody>
</table>

In a study where the Moderator, Theoretician, Summariser, and Source Searcher roles were compared to a baseline condition with no role (Schellens et al., 2007), only the Summariser role was seen to be of benefit in terms of an increase in the average level of knowledge construction; the Moderator and Theoretician roles were equivalent to being left without a role, while the Source Searcher role had a negative effect. It may be the case that participants felt constrained by the Source Searcher role, harming their performance (Wise et al., 2012; Fischer et al., 2013): they focused on offering new sources and did not feel free to build on what others already said. Another study (De Wever et al., 2010) also found that the Summariser role was associated with
5.1. Introduction

The single-duty roles are described in Table 5.1.

Table 5.2: Mapping from single-duty roles to the composite roles used in this thesis.

<table>
<thead>
<tr>
<th>Role</th>
<th>Practising Researcher</th>
<th>Research Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic Leader</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td>Moderator</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td>Source Searcher</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Theoretician</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Summariser</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

the greatest increase in knowledge construction, with Starter and Source Searcher showing the least benefit.

In contrast, later work on role assignment (Wise & Chiu, 2014) found that students who took on the role of Synthesizer or Wrapper (both sub-types of Summariser) during one weekly discussion went on to engage with the forum less frequently in subsequent weeks. While they were in role, their contributions to the discussion demonstrated greater depth than the other participants. However, in subsequent weeks, their participation in the task was on average lower in both quantity and quality than the rest of the participants; they were more likely to scan messages written by others, instead of reading them carefully, and less likely to reread their own messages. The Summariser role could thus be considered to have a negative effect over the longer term on those students who undertook it.

In the work described in this thesis, the two assigned roles (Practising Researcher and Research Expert) were composite roles (Gašević, Adesope, et al., 2015), made up of a combination of single-duty roles (Table 5.2). Participants in the Research Expert role produced and shared a recorded presentation on a research topic of their choice and then went on to lead the discussion on that topic. The other participants, in the Practising Researcher role, were required to ask questions and provide feedback. The effort required from participants when in the role of Research Expert was thus significantly greater than that in the Practising Researcher role. In addition, the Research Expert had overall responsibility for managing the discussion, in a similar way to those assigned a Summariser role in Wise and Chiu (2014).

In the study reported in the next section, we investigated how the order in which participants took on the two composite roles (Practising Researcher and Research Expert) affected the cognitive quality of their discussion contributions over time. We
used ENA to compare the mean cognitive quality of the messages sent by students in different groups, based on when they took on the *Research Expert* role. A similar approach was used to compare emergent roles and scripted roles in online discussions (M. Ferreira et al., 2021).

### 5.2 Peer-reviewed publication: Ordering Effects in a Role-Based Scaffolding Intervention for Asynchronous Online Discussion

This section includes the verbatim copy of the following peer-reviewed publication, reprinted with permission from Springer Nature:


**Contributions:** The ideas and analysis in the paper were developed and discussed between all authors of the work. The original idea, the experiments, and the writing were the work of the first author.
Ordering Effects in a Role-Based Scaffolding Intervention for Asynchronous Online Discussions

Elaine Farrow\textsuperscript{1}(\Letter), Johanna Moore\textsuperscript{1}, and Dragan Gašević\textsuperscript{2}

\textsuperscript{1} School of Informatics, University of Edinburgh, Edinburgh, UK
\{Elaine.Farrow, J.Moore\}@ed.ac.uk
\textsuperscript{2} Faculty of Information Technology, Monash University, Melbourne, Australia
Dragan.Gasevic@monash.edu

Abstract. A common scaffolding approach in computer-supported collaborative learning is the assignment of specific roles to the participants in online asynchronous discussions. Previous work has demonstrated how this type of scaffolding can result in student contributions of greater depth and quality. However, since students necessarily experience the roles in varying orders, it is important to consider whether the ordering impacts the outcome. This paper addresses the issue by examining a scaffolding intervention that was deployed in an asynchronous online discussion forum, where students were assigned to lead the discussion in one thread as the ‘expert’ and to participate in other threads by asking questions. A network analytic approach was used to visualise and quantify several potential ordering effects within the intervention. The constructs of cognitive presence and cognitive engagement, from the Community of Inquiry and the ICAP frameworks, were used together to measure the depth and quality of the discussion contribution expressed in each message. The analysis confirmed that the contributions made while the student was in the ‘expert’ role scored significantly higher for both constructs, but found that the order in which students took on each role had little impact on the quality of their contributions to other threads. This result contrasts with earlier work on single-duty roles that found an advantage in being assigned certain roles early in the discussion, and suggests that instructors should feel confident in rotating more complex user roles between students.

Keywords: Online discussion · Scaffolding · Critical thinking · Student engagement · Learning analytics · Community of inquiry · Cognitive presence · Cognitive engagement · Epistemic network analysis

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.

© Springer Nature Switzerland AG 2021
1 Introduction

Asynchronous online discussion forums are a common feature in virtual learning environments. Stand-alone discussion platforms such as Piazza\(^1\) are also used to manage students’ questions, in both online and classroom-based courses. The discussions that take place in these forums can help to build a sense of community among learners [6] – particularly important when in-person interaction was severely restricted due to a global pandemic. Time-stamped transcripts of the messages that are exchanged can also be used to inform research into how effective learning takes place through discussion.

Research in computer-supported collaborative learning (CSCL) has shown that participation in asynchronous discussions can be beneficial to participants, giving them opportunities to increase the depth of their own cognitive engagement through collaborative knowledge construction [12–14] as well as fostering social belonging [6]. However, in order to achieve these benefits, it is often necessary to provide explicit guidance in the form of scaffolding [11,15,18]. Prior work [26] suggests that when students are assigned a role that requires them to summarise the contributions of others, there is a positive effect on their breadth of listening while they are ‘in-role’, but the effect is not sustained afterwards. Other studies [7,21] have suggested that the timing of role assignment can impact outcomes, with earlier assignment seen as more beneficial for some roles.

The depth and quality of student participation in asynchronous online discussions has been examined and quantified using many different theoretical frameworks (e.g. Bloom’s Taxonomy [2], the SOLO taxonomy [1], The Community of Inquiry framework (CoI) [13,14], and the Interactive-Constructive-Active-Passive (ICAP) framework [3]). Of these, only CoI was designed specifically for the online context. Most previous studies have focused on a single framework, while a few have used a combination of two or more in order to provide a richer, multi-dimensional analysis of the data [8,9,19,22].

The specific type of scaffolding intervention considered in this work is an approach centred on assigning ‘roles’ to discussion participants. The study presented here investigated how the effect of the role-based scaffolding was moderated by the order in which participants experienced the different roles. Messages were classified using both the phases of cognitive presence from CoI and the modes of cognitive engagement in ICAP.

2 Background

2.1 The Community of Inquiry Framework

The Community of Inquiry (CoI) framework defines three ‘presences’ that support learning in an online community: social presence, teaching presence, and cognitive presence [14]. Of these, cognitive presence is considered to be the most fundamental to educational success. Discussions are expected to progress through

\(^1\) https://piazza.com.
its four phases (Triggering Event, Exploration, Integration, and Resolution) over time (Table 1), and the phases have been used as a measure of the depth and quality of student participation in asynchronous discussions [8,9]. In the ideal case, a discussion would start with a Triggering Event that defines the problem, move through an Exploration phase where new ideas are considered, then bring some of those ideas together in the Integration phase, and finally achieve consensus on a solution in the Resolution phase. In reality, progression through the phases is seldom linear. Many discussions do not reach the Resolution phase.

Table 1. The four CoI phases of cognitive presence in ascending order, plus the Other label, which can be used where a message does not display any cognitive presence.

<table>
<thead>
<tr>
<th>Short label</th>
<th>Phase of cognitive presence</th>
<th>Example behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIG</td>
<td>Triggering Event</td>
<td>Asking a question or posing a problem</td>
</tr>
<tr>
<td>EXP</td>
<td>Exploration</td>
<td>Exchanging ideas</td>
</tr>
<tr>
<td>INT</td>
<td>Integration</td>
<td>Integrating ideas and constructing meaning</td>
</tr>
<tr>
<td>RES</td>
<td>Resolution</td>
<td>Reaching consensus or suggesting a new hypothesis</td>
</tr>
<tr>
<td>OTH</td>
<td>Other</td>
<td>Commenting with no sign of cognitive presence</td>
</tr>
</tbody>
</table>

2.2 The ICAP Framework

The ICAP framework [4] has been used widely, in classroom-based studies as well as online. It defines four modes of cognitive engagement, based on observable student behaviours: Interactive, Constructive, Active, and Passive. Each mode represents a qualitatively different type of knowledge growth. Interventions and activities that targeted the higher modes of cognitive engagement were shown to achieve greater learning gains. Several recent studies adapted and expanded the original framework [8,9,25,28] in the context of asynchronous online discussions. The Constructive and Active modes were each subdivided and messages of Affirmation were treated separately (Table 2).

3 Research Question

Previous studies have shown how external scripts such as assigned roles can help students to develop skills relating to collaboration and social knowledge construction [7,11,21,24,27]. However, there is some evidence that the effects may not persist after the intervention has ended [26]. Some roles have been shown to be particularly beneficial to those who take them on (e.g. ‘summarizer’ [21]). Other single-duty roles have been shown to be detrimental to learning (e.g. ‘source-searcher’ [7]) when used in isolation. It is therefore seen to be important to rotate single-duty roles among students and to consider the use of composite roles that combine several lower-level duties [15,27]. The timing of role assignment has also been seen to impact learning outcomes [7]. There is thus a need for research into potential ordering effects within role-based interventions, since participants necessarily experience the roles in varying orders.
Table 2. The extended set of ICAP modes of cognitive engagement in descending order, plus the Off-task label, used for messages displaying no cognitive engagement.

<table>
<thead>
<tr>
<th>Short label</th>
<th>Mode of cognitive engagement</th>
<th>Example behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Interactive</td>
<td>As for C1, in response to earlier message content</td>
</tr>
<tr>
<td>C1</td>
<td>Constructive Reasoning</td>
<td>Displaying explanation or reasoning about the current topic</td>
</tr>
<tr>
<td>C2</td>
<td>Constructive Extending</td>
<td>Introducing new content to the discussion</td>
</tr>
<tr>
<td>F</td>
<td>Affirmation</td>
<td>Affirming what was said in an earlier message</td>
</tr>
<tr>
<td>A1</td>
<td>Active Targeted</td>
<td>Referencing specific previous content</td>
</tr>
<tr>
<td>A2</td>
<td>Active General</td>
<td>Showing other signs of being engaged with content</td>
</tr>
<tr>
<td>P</td>
<td>Passive</td>
<td>Reading messages without responding</td>
</tr>
<tr>
<td>O</td>
<td>Off-task</td>
<td>Commenting with no relation to the topic/course</td>
</tr>
</tbody>
</table>

Earlier studies looked at the effects of role assignments using a single measure of the quality of knowledge construction [21], sometimes in combination with final exam scores [7]. Recent work has shown the benefits of integrating insights from multiple frameworks for analysing aspects of student participation in asynchronous discussion tasks [8,9,19,22]. The research question addressed in the present study was therefore:

**RQ:** How do ordering effects between roles affect the depth and quality of student contributions to an asynchronous discussion task, as measured by both the CoI phases of cognitive presence and the ICAP modes of cognitive engagement?

4 Method

4.1 Description of the Data

The role-based scaffolding intervention examined in this study was deployed in a credit-bearing distance-learning course in Software Engineering over six course offerings (2008–2011). The discussion task accounted for 10% of the course grade and helped students to develop their own research questions. Two complex user roles were defined, with students expected to take on both roles during the task.

– **Research Expert:** prepare and upload a presentation about a relevant research paper of their choice, then lead a discussion on its content on a dedicated thread in the discussion forum; and
– **Practising Researcher:** contribute to discussions about other students’ presentation topics.

Both roles thus incorporated duties defined in earlier work as ‘summarizer’, ‘source searcher’, and ‘theoretician’ [7,21,27]. The Research Expert role additionally required the student to undertake ‘moderator’ and ‘topic leader’ duties.
The discussion task ran during weeks 3–6 of each course offering. Every student was expected to take on the *Research Expert* role once and the *Practising Researcher* role several times. Approximately one-third of the students took on the *Research Expert* role in each of the first three weeks of the task. The discussions that followed were asynchronous and ranged from 3 days to 27 days in duration. The median thread duration was 13 days. All discussion threads remained open until the end of the task. It was thus very common for a student to contribute to one or more threads as a *Practising Researcher* at the same time as they were acting as the *Research Expert* in their own thread.

In order to examine possible ordering effects in the present study, we considered two different ways of grouping messages by time, and another metric derived from those and intended to capture role order more directly (Table 3).

### Table 3. Labels assigned to messages in threads where students A, B, C, and D took on the *Research Expert* role in the first, second, second, and third week, respectively.

<table>
<thead>
<tr>
<th>Thread</th>
<th>Student</th>
<th>Thread Week</th>
<th>Expert Week</th>
<th>Role Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert-A</td>
<td>B</td>
<td>1</td>
<td>2</td>
<td><em>BeforeExpert</em></td>
</tr>
<tr>
<td>Expert-B</td>
<td>C</td>
<td>2</td>
<td>2</td>
<td><em>WhileExpert</em></td>
</tr>
<tr>
<td>Expert-D</td>
<td>A</td>
<td>3</td>
<td>1</td>
<td><em>AfterExpert</em></td>
</tr>
</tbody>
</table>

Our analysis focused on the messages sent by students while they were in the *Practising Researcher* role, for two reasons: we wanted to distinguish potential role ordering effects from the large effect of the role assignment intervention itself [9,15]; and each student was only the *Research Expert* once. We excluded 9 messages that were sent by participants who never took on the *Research Expert* role, leaving 891 messages from 84 threads (Tables 4 and 5).

Each message was assigned one label from each theoretical framework, based on its textual content. Two expert coders labelled the messages with the CoI phases of cognitive presence (Table 1), achieving high levels of agreement (98.1%...
Table 4. Counts of unique participants, threads, and messages.

<table>
<thead>
<tr>
<th>Thread</th>
<th>Week</th>
<th>Expert Week</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week</td>
<td>Threads</td>
<td>Messages</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>352</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>288</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>251</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>891</td>
</tr>
</tbody>
</table>

Table 5. Message counts in the Role Order groups.

<table>
<thead>
<tr>
<th></th>
<th>BeforeExpert</th>
<th>WhileExpert</th>
<th>AfterExpert</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Messages</td>
<td>310</td>
<td>304</td>
<td>277</td>
<td>891</td>
</tr>
</tbody>
</table>

agreement, Cohen’s $\kappa = 0.974$). A second pair of independent coders assigned labels from the extended set of ICAP modes of cognitive engagement (Table 2), achieving ‘substantial’ inter-annotator agreement (Cohen’s $\kappa = 0.623$) [17].

For the purposes of the present study, the Interactive and Constructive Reasoning labels were combined together and only Constructive Reasoning was used. The primary difference between them is that a message can only be labelled as Interactive if it is a direct response to the substantive content of a previous message (Table 2). While Interactive messages were relatively common for a Research Expert, there were limited opportunities for a Practising Researcher to interact in such a way during the discussion task presented in this study.

4.2 Epistemic Network Analysis

Epistemic Network Analysis (ENA) [23] is a network analytic approach that is designed for analysing the connections between small sets of concepts in a densely connected network. It allows sub groups to be compared both visually and statistically, and has been widely used in studies of online discussions [16,29] in general, and specifically for the constructs of cognitive presence and social presence in a Community of Inquiry [10,19,20].

Co-occurrences of labels in the data are used to construct a high-dimensional concept network. The conversation parameter defines which connections are included in the analysis. The network is projected down onto the two most informative dimensions, while maintaining the mathematical relationships between concepts, using singular value decomposition. The relative positions of the concept nodes in the resulting projection space makes the space itself interpretable, because concepts that share a pattern of connections will tend to be located close together [23]. A single point in the projection space represents the weighted mean of the connections in one sub-network, defined by the unit of analysis parameter. For example, this could be all the messages in a thread.
In the present study, we set the conversation parameter to be a single message. This meant that the only connections included in the network were the pairs of labels from the two theoretical frameworks: one label from the CoI phases of cognitive presence and one from the ICAP modes of cognitive engagement. We first grouped the messages by student and thread, so that all the messages sent by the same student in a single thread were aggregated together. As we looked at the order-based groupings in turn, the messages were aggregated further.

To become familiar with the general associations between the individual CoI phases of cognitive presence and the ICAP modes of cognitive engagement in this data set, we first explored the overall mean network based on all the messages sent in both roles. Noting the locations of the nodes in the overall mean network allowed us to interpret the space in terms of the framework constructs. The same projection space was subsequently reused for the analyses of messages sent by Practising Researchers, broken down by each of the different groupings (Thread Week, Expert Week, and Role Order). The messages were aggregated by student, thread, and group to create the data points for each network. We used Mann-Whitney tests to determine whether pairs of groups were significantly different along either of the two axes of the projection space.

5 Results

Figure 1 shows the average ENA network across all messages. The framework constructs are shown using their short labels (Tables 1 and 2) to reduce visual clutter. The X axis accounts for 21.7% of the variance in the data and the Y axis accounts for 20.2%. The X axis primarily distinguishes between the early phases (Triggering Event and Exploration) and the later (Integration) phase of cognitive presence, while the Y axis distinguishes linearly between the three highest ICAP modes of cognitive engagement. The direct effect of the role assignment intervention is clearly visible. The points representing messages sent by students in the Research Expert role are all found toward the upper left of the plot, in the vicinity of the Constructive Reasoning (C1) node. In contrast, the messages sent by those in the Practising Researcher role are dispersed throughout the projection space, with the group mean near the centre of the plot.

Figure 2 shows the projection networks comparing messages sent by Practising Researchers, aggregated by Thread Week, Expert Week, and Role Order. These networks all use the same projection space as Fig. 1. The axes account for slightly less of the variance in the data: 21.1% for the X axis, and 20.1% for the Y axis. In each case, the group means appear close together, indicating that any differences are small. A series of Mann-Whitney tests showed that there were no statistically significant differences at the $\alpha = 0.05$ level between any of the Thread Week values in Fig. 2(a). In addition, ExpertWeek1 and ExpertWeek2 were not significantly different from each other in Fig. 2(b). The small difference seen between ExpertWeek2 and ExpertWeek3 along the Y axis was not considered significant after Bonferroni correction. However, ExpertWeek3 was significantly different at the $\alpha = 0.05$ level along the Y axis (V2) from ExpertWeek1 ($U = 24809.00, p = 0.0007, r = 0.18$). This indicates that students who
were in the last group to take on the Research Expert role tended to contribute to other threads at a lower level, as measured by the ICAP modes of cognitive engagement, compared to their counterparts in the first group. The effect size is small [5]. There was no significant difference along the X axis.

Considering the effect of Role Order in Fig. 2(c), a series of Mann-Whitney tests confirmed that, after Bonferroni correction, the only statistically significant difference between the groups was between the AfterExpert group and the the BeforeExpert group along the Y axis ($U = 29967.00$, $p = 0.0094$, $r = 0.13$). This indicates a small effect size for Role Order, corresponding to a tendency for students to demonstrate higher levels of the ICAP modes of cognitive engagement in threads that were started in the week(s) after their own expert thread, compared with the threads that were started in the week(s) before their own.
Fig. 2. ENA networks constructed using the messages sent in the Practising Researcher role, grouped by (a) Thread Week, (b) Expert Week, and (c) Role Order, in addition to student and thread. The points are coloured according to the relevant group: (a) the week within the task when the thread started; (b) the week within the task when the message author started acting as Research Expert; (c) the Role Order label. In both (a) and (b), week 1 is shown in red, week 2 in blue, and week 3 in purple. In (c), BeforeExpert is shown in blue, WhileExpert is in red, and AfterExpert is in purple. Group means are labelled and shown as squares in the appropriate colour. (Color figure online)

6 Discussion

The results of this study confirmed the ability of a role-based scaffolding intervention to effect positive change, as seen in previous work [9,15] where messages sent by students in the Research Expert role achieved greater depth on both the CoI and ICAP frameworks, compared with those sent by students in the Practising Researcher role. However, there was little evidence of ordering effects. No significant differences were found between the message threads that were started in the first week of the activity compared to those in the final batch, despite the much longer time available for students to develop a deeper discussion.
The students in the present study were always assigned to a role, and these were composite roles that incorporated several of the low-level single-duty roles investigated in previous work. We noted a small effect where the group of students who were last to take on the Research Expert role demonstrated lower cognitive engagement in their contributions to other threads. One explanation for this could be that the effort of leading their own thread in the final week, while also ensuring that they had fulfilled the participation requirements, led to shallower engagement. Another potential explanation is the timing effect found in prior work [7], where a cohort that began without roles and had them assigned later performed worse than a cohort that had roles from the beginning. It is possible that the group who took on the Research Expert role last did not fully engage with the Practising Researcher role earlier.

A small effect was also found in the analysis of Role Order. Students in the Practising Researcher role, contributing to threads that started in the weeks after their own Research Expert thread started, scored higher on the ICAP modes of cognitive engagement. This could be because those students had time to devote to asking deeper questions, having finished with their own presentation. It could also be because they had learned from the experience of being in the Research Expert role and used this knowledge in later situations [7].

Since the discussion task in the present study only ran for four weeks, we were not able to discover any longer-term effects on behaviour. Analysis of discussions that took place over a longer period could produce different results, as participants grow in confidence and develop their skills, or perhaps become disengaged. The nature of the discussion task meant that students were often managing both roles in parallel: leading their own thread as a Research Expert, while at the same time contributing to other threads as a Practising Researcher. More specific instructions were given to participants in the later course offerings regarding the minimum contribution expected from students in the Practising Researcher role. The present study did not distinguish between those cases.

7 Conclusion

In the role-based scaffolding intervention presented in this study, the effects of role order were found to be small – especially in the context of the large primary effect of the intervention in improving student contributions according to two separate measures of depth and quality. This result suggests that instructors should feel confident in assigning complex roles and rotating them between students, without being afraid that a particular ordering might lead to disadvantage. Since the discussion task in the present study was relatively short in duration, future work should look at behaviour over the longer term, and in particular at examples where students repeat a similar style of task over time. It would also be valuable to directly contrast the use of single-duty roles with composite roles like those used here.
References


5.3 Summary of contributions

The network analytic approach presented in this chapter allowed the two frameworks that comprise cognitive quality to be examined together in the context of the role assignment intervention. The primary difference between the roles can be clearly seen in Figure 1 of Farrow et al. (2021b), reproduced on p. 114: while participants were in the Research Expert role, their contributions tended on average towards the higher indicators of quality on both frameworks (top left of the plot). There was much more variation in the quality of messages sent by those in the Practising Researcher role. The additional ENA networks that we constructed in order to compare groups based on role order (Figure 2 in Farrow et al. (2021b), reproduced on p. 115) allowed us to determine that most of the between-group differences were not significant.

The present study found no adverse effects from the practice of rotating roles among participants. This marked a contrast with earlier work (Wise & Chiu, 2014), which found a detrimental effect of role assignment on the students who took on a Summariser role, whereby the quality of their later discussion contributions was reduced. In fact, the present work found a small positive effect: once participants had experience of the Research Expert role, they demonstrated higher cognitive engagement in the contributions they made to threads that started in later weeks.

One major difference between the earlier study (Wise & Chiu, 2014) and the study described in the previous section of this thesis has already been discussed in the present chapter: the Practising Researcher and Research Expert roles were both composite roles, going beyond what was required of the Summariser single-duty role. In addition, students in the Research Expert role entered the discussion with prior knowledge of the topic, gained during the preparation of their recorded presentation. As a consequence, it was likely that they were better able to encode new information about the topic, encountered during the course of the discussion, compared to those in the Practising Researcher role (O'Donnell & Hmelo-Silver, 2013).

A second difference related to the sequencing of the roles and of the discussions. In the earlier study (Wise & Chiu, 2014), each discussion took place in a different week and every member of the group was expected to contribute to every discussion. In contrast, in the study reported in this chapter, many different discussion threads developed concurrently during each of the four weeks of the task. Participants could choose to contribute to some threads and not others, and it was typically the case that participants were managing both the Practising Researcher and Research Expert roles
in parallel. The task design meant that participants did not switch abruptly between roles from one week to another, as was typical in previous studies (Schellens et al., 2007; Wise et al., 2012; Wise & Chiu, 2014). In fact, there was no defined end point to their time as Research Expert – other participants could ask questions and continue the discussion at any time during the remaining weeks of the discussion task.

The temporal overlap in roles seen in the present study could partly explain why the cognitive quality of participant contributions did not decline after they had taken on the Research Expert role, as might have been expected based on prior work (Wise & Chiu, 2014). The increased sense of responsibility related to leading a discussion thread did not appear to diminish until the whole discussion task ended; the participants did not have the opportunity to disengage in later weeks. Further work is needed to explore possible longer-term effects of role allocation and ordering that might become evident in later tasks.
Chapter 6

Conclusions and future directions

Understanding the factors that influence the cognitive quality of student contributions to asynchronous online discussions is important because it can enable students and instructors to adopt practices that enhance quality and encourage deeper learning. The previous chapters of this thesis examined how measures of cognitive quality – and the two well-established frameworks that underpin it, CoI and ICAP – varied with the linguistic and structural attributes of the dialogue, and with the two instructional interventions that took place during the course from which our data was collected. In the present chapter, we consolidate and summarise the main contributions of this thesis in the context of the overarching research goal and the detailed research questions presented in Section 1.1. We go on to highlight the implications of our work for the field, and conclude by suggesting some possibilities for future work in this area.

6.1 Summary of contributions

The overall aim of this thesis was to model the cognitive quality of the contributions students make to asynchronous online discussions. In the work reported in Chapter 3, we presented empirical evidence that the constructs of two pre-existing and well-supported frameworks, CoI and ICAP, were generally aligned but not closely related, which meant that they could be used together to form a rich, two-dimensional measure of cognitive quality. We uncovered correlations between specific linguistic and structural attributes of the dialogue and the indicators of cognitive quality; for example, the number of words in a message is positively correlated with cognitive quality, while the number of question marks is negatively correlated (Table 3.1). We were surprised to discover that the order in which messages were added to the discussion was not predictive of
quality, even though messages that were nested more deeply within threads tended to be consistently higher quality. We had expected that the quality of messages added later in the discussion would tend to be higher than early replies, since their authors had greater opportunity to read and reflect, and a richer history (Suthers et al., 2010) to build on. Instead, messages added later in time at a shallow level of nesting were of similar quality to other shallow-nested messages, possibly indicating that messages in the other branches of the thread had not been read. In Chapter 4, we demonstrated that the external facilitation intervention, where the Treatment group received additional participation guidelines to support their self-regulation, changed the pattern of associations between the constructs of the two frameworks. For example, participants who sent messages demonstrating explanation and reasoning (Constructive Reasoning) tended to reach the later phases of cognitive presence (Integration and Resolution) more often if they were in the Treatment group rather than the Control group. This difference corresponds to a shift away from exploring a single idea in isolation, towards connecting ideas and building on what others have contributed. We extended a previous result (Gašević, Adesope, et al., 2015) by showing that the messages sent by participants in the role of Research Expert, compared to those sent while in the Practising Researcher role, exhibited greater cognitive quality in terms of the ICAP modes of cognitive engagement as well as the CoI phases of cognitive presence. In Chapter 5, we showed that the order in which discussion participants took on each of the two assigned roles across the weeks of the task had minimal impact on the quality of their contributions while in the Practising Researcher role, confirming that the chosen role rotation did not disadvantage any of the groups. The findings in this thesis demonstrate the benefits of using a multi-dimensional quality measure, and also indicate some practical steps that instructors can adopt to increase the cognitive quality of contributions to asynchronous online discussions.

In addition to our core findings, summarised in the previous paragraph, we also developed a practical approach (Chapter 2) to address a common issue that hinders reproducible research in this area: often, messages from discussion forums cannot be made available for research because the message content includes personally identifying information. In order to better support such research, we developed a novel semi-automated approach that can be used to replace personal names in discussion forum messages with consistent pseudonyms. In our evaluation, we demonstrated the importance of handling nicknames and misspellings, which made up nearly one quarter of all the personal names used in our data set. This work was published in
6.2 Impact of the present work

6.2.1 Associations between theorised properties of cognitive quality and attributes of the dialogue (RQ1)

In Chapter 3, we extended previous work that had found correlations between particular linguistic and structural dialogue attributes in asynchronous online discussions and the CoI phases of cognitive presence (Kovanović et al., 2016; Neto et al., 2018; Farrow et al., 2019; Barbosa et al., 2020). Our work additionally identified correlations between the ICAP modes of cognitive engagement and dialogue attributes. This work is published in Farrow et al. (2022). The models that were trained to label the ICAP modes of cognitive engagement achieved notably higher Cohen’s $\kappa$ scores than those for the CoI phases of cognitive presence. We hypothesised that the ICAP modes are more closely related to the superficial characteristics of the message text than are the CoI phases of cognitive presence, captured in the linguistic features used in the models. Prior work (Rosé et al., 2008) has emphasised the importance of selecting model features that align with the design of the coding scheme.

We found that many of the same attributes that had been found to be predictive for the CoI phases of cognitive presence in earlier work were also predictive for the ICAP modes of cognitive engagement in our data set (Table 3.1). Such dialogue attributes can therefore be considered to be predictive of cognitive quality. We also hypothesised that the most informative dialogue attributes could be used directly as proxy indicators of cognitive quality, in situations where labelling messages with framework constructs is impractical. Of course, correlation is not causation; if participants changed their behaviour with the intention of changing such attributes, the cognitive quality of their contributions might not necessarily improve as a result (Knight et al., 2014). Further work is needed to “close the loop” (Koedinger et al., 2013) and test these ideas in practice.

Farrow et al. (2023). The code is available on github\footnote{https://github.com/efarrow/nicknames} and is expected to be applicable to discussions in other informal settings.

In the next section, we address the implications of our work for research and practice in the context of the research questions posed in Section 1.1.
In many earlier studies, message length was found to be one of the most predictive attributes for cognitive presence (Kovanović et al., 2016; Neto et al., 2018; Farrow et al., 2019; Barbosa et al., 2020), cognitive engagement (Yoge et al., 2018), and course performance (Wise & Cui, 2018). In our studies, described in Chapter 3, we again found that message length was highly predictive for both cognitive presence and cognitive engagement for the messages in our data set, with longer messages tending to be higher quality (Table 3.1). However, looking at some of the other linguistic attributes that were used as model features – for example, the number of question marks and the number of positive emotion words – we note that those counts tend to be correlated with message length. It is thus possible that the strong predictive effect of overall message length could mask the influence of a particular class of word when counts are used directly as model features. Future work should consider normalising such counts relative to the number of words in the message, to separate out the effects of message length from those of message content.

Similar concerns arise in relation to measures of lexical diversity: those that rely on type-token ratios (TTR) are known to vary as a function of message length, reducing their usefulness, with very short messages being assigned artificially high diversity scores (Yang et al., 2022). Vocabulary Diversity (VOCD) was designed to be an alternative measure of lexical diversity that could compensate for differences in text lengths. In our work, both TTR and VOCD lexical diversity measures were found to be highly predictive of cognitive quality (Table 3.1), but the direction of the relationship was reversed. Only the VOCD measure correctly linked higher lexical diversity with higher cognitive quality.

Our most notable and actionable finding in response to RQ1 was that the depth of a message within a discussion thread was positively correlated with cognitive quality, whereas the temporal order in which messages were posted was not predictive. With this in mind, we recommended rewarding participants for extending existing discussion threads, rather than adding a new top-level comment to the original message. In our data, many top-level comments were simple questions that did not add any new content to the discussion; for example, “Do you know of any other system which tags video with which to compare X to?” In contrast, in order to continue an existing discussion thread successfully, participants needed to read and engage with contributions from others (Wise et al., 2014), reflect and assess (Garrison et al., 2001), and then share information (Radkowitsch et al., 2020), challenge, clarify, or otherwise build on the work of their peers (Garrison, 2011; Chi & Wylie, 2014; Kimmerle et al., 2021). Similar
benefits might perhaps also be seen if participants worked together to edit a single shared resource on each topic, using a suitable collaborative platform such as a wiki (Scardamalia & Bereiter, 2021), rather than adding their thoughts sequentially in distinct messages (Chen et al., 2021). In such a scenario, it would again be vital to read and evaluate what had already been written before contributing new content. Future studies could compare these two approaches, assigning one group to edit a single shared document, while another group engaged in sequential discussion using a threaded discussion forum.

6.2.2 Associations between theorised properties of cognitive quality under different conditions (RQ2)

Our second research question asked about the associations between the CoI phases of cognitive presence and the ICAP modes of cognitive engagement, and the robustness of those associations under different instructional scaffolding conditions. Whenever multiple frameworks are developed independently to address similar concepts, such as the cognitive aspects of learning, it is important to consider how those frameworks might relate to one another. In the case of the CoI and ICAP frameworks, the results of the second study reported in Chapter 3 indicated that they generally provided complementary perspectives on cognitive quality in online discussions.

We considered that different instructional scaffolding conditions might alter the associations between the constructs of CoI and ICAP. Previous work had used ENA to compare cognitive and social presence within the CoI framework (Rolim, Ferreira, et al., 2019), but our study, reported in Chapter 4, was the first to use a network analytic approach to examine the associations between the CoI phases of cognitive presence and the constructs of a second theoretical framework – in our case, ICAP. We used cross-tabulations, heat maps, and ENA to examine how the labels from each framework co-occurred in our data set, and how those associations were moderated by two instructional interventions. The analytic approach we developed allowed insights from both frameworks to be integrated, and we expect it can be useful for research in other contexts where multiple relevant indicators have not previously been examined together. This work is published in Farrow et al. (2021a).

Two different instructional interventions were represented in the data we used. First, a role assignment intervention, where every participant took a turn of being the Research Expert, leading a discussion on a topic of their choice, and also contributed to
other discussions in the role of Practising Researcher. Second, an external facilitation intervention, designed with the aim of encouraging greater self-regulation, and involving a change in guidance between different sessions of the course. Although the external facilitation intervention directly addressed the behaviour of participants only while they were in the Practising Researcher role, it had a positive effect on the cognitive quality demonstrated in the threads as a whole. The additional guidance given to the Treatment group encouraged those in the Practising Researcher role to move beyond questions that explored a single idea, to make connections with external information sources and with other parts of the course. This positive change in participant behaviour was visible in the ENA networks that were constructed from the messages in the Control and Treatment conditions (Figure 6 in Farrow et al. (2021a), reproduced on p. 96). In turn, the higher quality contributions from Practising Researchers meant that participants in the Research Expert role were also able to contribute at a deeper level, building on their peers’ contributions and reaching the Integration and Resolution phases more often. On the basis of this finding, in response to RQ2, we conclude that instructors should not be afraid to focus an intervention on a subset of participants whose contributions are low quality. Raising the standard of the lowest-performing group may well have positive effects on the whole community (Zhang et al., 2019).

A second finding in response to RQ2 related to the Exploration phase of cognitive presence. Earlier work found the Exploration phase to be the most common (Kovanović et al., 2014; Hu et al., 2020; Hu et al., 2022b). In fact, participants can often become “entrenched” in that phase (Garrison, 2011, p. 47), needing external help in order to move on from Exploration to Integration. In our data set, the Exploration phase was the largest category, accounting for nearly 40% of all messages. The analysis in Chapter 4 looked at the ICAP modes of cognitive engagement assigned to those same messages. In the data set as a whole, Exploration messages were distributed fairly evenly across the Active Targeted, Constructive Extending, Constructive Reasoning, and Interactive modes. However, differences were evident between the various conditions. Many of the Exploration messages sent by participants in the Practising Researcher role referenced specific previous content but did not add anything new (Active Targeted mode). In contrast, the majority of the Exploration messages sent by those in the Research Expert role introduced new information to the discussion (Constructive Extending mode), or demonstrated explanation and reasoning (Interactive mode). Our research in response to RQ2 therefore shows how the large class of Exploration messages can be subdivided in a theoretically meaningful way by using the ICAP modes of cognitive engagement.
A further actionable finding of the study reported in Chapter 4 is that it may be beneficial for researchers to combine the Constructive Reasoning and Interactive modes together for analysis, as some have already done (Wang et al., 2016b; Vellukunnel et al., 2017), and as we did in Chapter 5. This recommendation is based on the definition of Interactive mode (Table 2.2), which specifies that the message must be a reply to a previous Constructive message. By definition, the initial message in a thread can never gain that label; and, unless the initial message contains substantive content, none of its direct replies can be labelled Interactive either. Although it is often better to reuse existing labels than to invent new ones (De Wever et al., 2006), a case could be made for amending the extended ICAP taxonomy further. In Chapter 2, we introduced the Affirmation label in order to allow messages of thanks and agreement to be labelled independently, so that a later revision to the label of a parent message would not affect the label assigned to the child. In the same way, it might be sensible to introduce a new label to identify Constructive Reasoning messages that are direct responses to an earlier message; for example, Constructive Reply. The new label could be used regardless of the label assigned to the parent message. When a message constitutes a direct response to an earlier message in this way, the relationship between them can be referred to as uptake (Suthers et al., 2010; Kimmerle et al., 2021). If a Constructive Reply message did in fact build on reasoning in the earlier message – making it eligible to be relabelled as Interactive during an optional post-processing step – that would also be evidence of transactivity (Suthers et al., 2013). Both uptake and transactivity are central concepts in the study of collaborative learning within CSCL.

6.2.3 Effects of temporal ordering (RQ3)

Our final research question addressed the issue of role rotation and potential effects caused by the order in which participants experience the roles. In previous studies of role assignment, there were cases where taking on a particular role was beneficial at the time but harmful overall (Wise & Chiu, 2014), or where those who experienced the roles in a particular order benefited more than others (De Wever et al., 2010). While the role assignment intervention we examined in Chapter 4 was found to be effective in raising cognitive quality for those in the Research Expert role, compared to their contributions in the Practising Researcher role, that study did not consider the temporal ordering of the roles. In Chapter 5, we focused on the role assignment intervention week by week, Constraints based on thread position were addressed in earlier CSCL work, where thread-initial messages were constrained to be externalizations (Rosé et al., 2008).
in order to determine whether the order in which participants experienced the different roles affected the cognitive quality of their contributions to the discussion. This work is published in Farrow et al. (2021b).

The network analytic approach we adopted allowed us to compare the behaviour of participants across the weeks of the discussion task as they took on each of the roles. Our analysis found no adverse effects related to role order, while confirming the large positive primary effect of the role assignment intervention. It is important to note that the two roles in our study, Practising Researcher and Research Expert, were composite roles encompassing several duties (Gašević, Adesope, et al., 2015), unlike those in earlier studies, which involved only a single duty (Strijbos & Weinberger, 2010). From a practical perspective, our results mean that instructors can feel confident in assigning composite roles in rotation.

Future work could use the same data set to examine in more detail how the different threads were interleaved in time. During the labelling process, we found very few explicit references to content discussed in a different thread, so we felt confident in labelling each thread independently with the ICAP modes of cognitive engagement. However, it is possible that the extent and quality of responses from participants to each thread may have been affected by the other threads in which they were engaged concurrently, as well as by their assigned role (Practising Researcher or Research Expert). For example, a participant might choose not to contribute to a newly started thread because they were already engaged in discussions in several other threads. The formation of social ties in the form of reciprocal commenting behaviour (Joksimović et al., 2019) could perhaps also be influenced by the role rotation. We did not look for evidence of such patterns of behaviour. Additionally, just as prior work on uptake analysis (Suthers et al., 2010) incorporated interactions with content from other threads into contingency graphs, from which uptake graphs can be derived (Suthers et al., 2010; Trausan-Matu & Slotta, 2021), it could be informative to track the cognitive quality of the contributions from each participant across multiple threads to look for more complex interactions as the conversations evolve over time.

6.3 Future directions

There are many possible future directions for our work and we have indicated some ideas for potential studies in the previous section. In addition, there are two further strands of research we can envisage. The first strand would extend the approach and
6.3. Future directions

The data set we used throughout this thesis came from a fully-online distance-learning course at Masters level. It would be valuable to explore how our findings translate to other settings. Undergraduate and non credit-bearing courses might show different results, influenced by the maturity and motivations of participants. Fully-online courses are not the only ones to incorporate discussion forums. Traditional face-to-face courses are increasingly embracing the use of online communication platforms, allowing discussions within large classes of students to be managed by teams of tutors. The Covid-19 pandemic led to many abrupt changes to the education landscape, including greater use of online platforms for communication between students and instructors. It is expected that some of those changes will remain in place, even as restrictions are lifted. If so, such platforms could offer new opportunities to investigate the role of asynchronous online discussions within otherwise traditional courses. The importance of building and maintaining social connections gained greater attention while people were not able to gather in groups, and future work should certainly incorporate consideration of social presence alongside cognitive presence.

Recent technological developments also suggest new possibilities for related research. Neural network (NN) models have led to major advances in many areas in recent years. NNs remove or reduce the need for researchers to extract linguistic features from the text; the models themselves are able to learn from patterns in the training data in a flexible way, going far beyond the “rigid rules” (Rosé et al., 2008, p. 246) of earlier approaches and taking proper account of the context in which words are used. Novel NN approaches to handling text, such as word embeddings (Mikolov et al., 2013; Pennington et al., 2014) and transformers (Devlin et al., 2018) have the potential to be useful tools for learning analytics (Atapattu et al., 2019; Hu et al., 2022a; Lee et al., 2022).

The random forest classifiers we used in Chapter 3 are relatively simple in comparison to NNs, but have the advantage of interpretability. While NN models may achieve greater predictive accuracy than other classifiers, there is no guarantee that the features they learn from the raw text data will be meaningful to researchers, practitioners, or students. NN models may also introduce new biases into the modelling process (Oshima & Hoppe, 2021). However, Explainable AI is a growing field and many approaches are under development to make NN models more understandable (Hu, Ferreira Mello, et al., 2021), thereby enabling students and practitioners to benefit from them while retaining agency (Wise et al., 2021). Future work should investigate
how well NN models can assign labels corresponding to the CoI phases of cognitive presence and the ICAP modes of cognitive engagement, based on the text content of the messages, with or without structural and other contextual information. If such a model could label new messages quickly and reliably, it would be of practical use even if its mechanisms remained opaque.

One particular advantage of NNs is that their internal representations are sometimes language independent (K et al., 2020). A word embedding converts textual input into vectors in a high-dimensional space, within which many languages can be represented simultaneously. The rest of the model is then language-agnostic. Such a model could be trained on messages in one language and then deployed to label messages in another language, benefiting language communities that do not have access to NLP tools like LIWC and Coh-Metrix (Barbosa et al., 2020). Questions about cultural differences in language use (Borge & Rosé, 2021) could then be addressed in new ways. Other NN techniques, such as multi-task and transfer learning (Collobert & Weston, 2008; Hosmer & Lee, 2021), could allow both CoI and ICAP labels to be assigned by the same model, allowing future researchers to make use of multi-dimensional quality measures more easily.
Appendix A  Extended ICAP Coding Manual

Annotating discussions centred on an unseen artefact using an extended version of the ICAP framework

Elaine Farrow
2 September 2020

1 Introduction

The annotation guidelines presented in this document were developed to inform the labelling of student discussion forum messages using an extended version of the ICAP framework (Chi and Wylie, 2014), particularly in the context of an artefact, such as a video presentation, that is not available to the annotators. The extended ICAP framework has been used to label messages from MOOC discussion forums and annotated textbooks (Wang et al., 2016b; Yogev et al., 2018). The guidelines in that earlier work were adapted here to allow the framework labels to be applied to a discussion where each message thread is a response to a video presentation that is not available to the annotators (Farrow et al., 2020).

The ICAP framework defines cognitive engagement based on observable behaviours. It looks at individual learning activities and how they relate to students’ cognitive engagement with the learning materials. Four modes of cognitive engagement are identified, and the framework predicts that higher modes will be correlated with greater learning gains. In descending order, these modes are Interactive, Constructive, Active, and Passive. Each mode 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 1). Off-task behaviours do not constitute any sort of cognitive engagement.

The lowest mode in the framework, Passive engagement, corresponds to the least taxing on-task activities; for example, watching a video or reading a discussion forum post. Active engagement covers activities that demand the student’s attention, such as taking verbatim notes or reworking previous content. To qualify as Constructive engagement, novel output must be generated; for example, summary notes that link together concepts, or a list of relevant external resources. Interactive engagement requires interaction with a partner, and both partners must be engaged constructively. Off-task behaviours do not constitute cognitive engagement at all.

Prior work has demonstrated the feasibility of applying a modified version of the ICAP framework to MOOC discussion forums (Wang et al., 2016b) and to student comments on an annotated electronic course text (Yogev et al., 2018). In common with Yogev et al. (2018), this work looks at the case where the discussion centres around a pre-existing artefact. In the earlier work, that artefact was an electronic course text. Participants highlighted sections of
the text and added their own comments, questions, and answers, such that the discussion was grounded in the text. Here, we present guidelines for assigning labels to a data set where the relevant artefact is a video presentation, prepared by one of the discussion participants, and not available to the annotators. Annotators must therefore judge based on context whether a comment is simply paraphrasing, repeating, or requesting clarification about content from the recording (Active mode); or whether it introduces something new beyond what was already given (Constructive mode).

2 The extended label set

We build on the extended cognitive engagement taxonomy used in Yogev et al. (2018). This allows for finer-grained distinctions between messages within two of the original modes: Constructive mode is divided into Constructive Reasoning and Constructive Extending, while Active mode is divided into Active Targeted and Active General.

Like Yogev et al., we also treat affirmation messages as a special case. However, we differ in the way we handle these messages during the labelling process. In that earlier work, messages of agreement or thanks expressed in response to an earlier message had a label assigned to them that depended on the label of that earlier message. 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.

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 helpful reply”), and appearing in the same position within a discussion thread, could receive different labels depending on the labels of the earlier messages. Therefore, our guidelines do not assign the derived label to affirmation messages directly. Instead, the Affirmation label is used as a placeholder for messages that affirm what another user said about the current topic in an earlier message (e.g. an agreement or thank you). Once all the messages in the data set
have had labels assigned (manually 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.\footnote{See Farrow et al. (2020) for further discussion on this point.}

The full extended ICAP label set is presented in Table 1. All of the labels are described in more detail in Section 4, and positive and negative examples are given for each label based on messages exchanged in an online Software Engineering course where students were responding to video presentations.

<table>
<thead>
<tr>
<th>Cognitive engagement mode</th>
<th>Label</th>
<th>Example behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Off-task</td>
<td>O</td>
<td>Commenting without any relation to the current topic or the course content</td>
</tr>
<tr>
<td>Interactive</td>
<td>I</td>
<td>Displaying explanation or reasoning about the current topic in response to an earlier message</td>
</tr>
<tr>
<td>Constructive Reasoning</td>
<td>C1</td>
<td>Displaying explanation or reasoning about the current topic</td>
</tr>
<tr>
<td>Constructive Extending</td>
<td>C2</td>
<td>Introducing new content about the current topic to the discussion</td>
</tr>
<tr>
<td>Affirmation</td>
<td>F</td>
<td>Affirming what was said in an earlier message</td>
</tr>
<tr>
<td>Active Targeted</td>
<td>A1</td>
<td>Referencing specific previous content</td>
</tr>
<tr>
<td>Active General</td>
<td>A2</td>
<td>Showing other signs of being engaged with course content</td>
</tr>
<tr>
<td>Passive</td>
<td>P</td>
<td>Reading messages without responding</td>
</tr>
</tbody>
</table>

3 The labelling process

Every message is assigned a single label: either the Off-task label, or else a specific label from the extended ICAP label set. On-task messages can relate to the current topic for a particular discussion thread, or can address the course content more generally (see definitions below). A message that does not relate to either course content or the current topic is Off-task. The definitions of current topic and course content presented below are also used in the label definitions for on-task messages in Section 4.

If a message includes evidence of multiple on-task behaviours, the label corresponding to the highest mode of cognitive engagement that was identified in the message is chosen. For example, if a message demonstrates both Interactive and Affirmation behaviours, it is labelled as Interactive.
3.1 Current topic

The current topic is defined by the content of the video presentation, where a research paper is presented. The current topic includes the academic content discussed in the presentation; the concepts, frameworks, tools, experiments, results, and theories described in the paper; and the paper itself as an object of study.

However, technical and stylistic aspects of the video presentation – such as comments on presentation skills, the clarity of the audio, or the design of the slides – are not considered part of the current topic; they belong, instead, to the broader course content.

3.2 Course content

Course content includes course logistics; lectures, presentations, reading lists, and assignments; and relevant personal experience. Course content does not include personal introductions or motivation for taking the course (these are considered off-topic).

4 Label definitions and constraints

There are two general constraints on the use of certain labels. The first is that messages can only be labelled as Interactive or Affirmation if they are responding to the content of a previous message in the discussion – to build on it or to affirm it, respectively. In particular, this means that where a message simply presents a link to a video presentation or a research paper and has no substantive textual content itself, the replies to that message cannot be labelled as Interactive or Affirmation.

The second important constraint is that Interactive and Constructive messages must relate to the current topic (Section 3.1), and not only to general course content (Section 3.2). In this, we follow Wang et al. (2016a), who say, “If the student is reasoning about scoring rubric, that’s categorized as active behavior, because that only shows the student is paying attention to course materials, but not that he/she is engaged in constructive behavior for the purpose of learning.”

In the rest of this section, we define each of the labels and give positive and negative examples from an English language data set to illustrate the definitions and constraints.

4.1 Off-task messages (O)

Off-task messages only talk about content that is totally unrelated to both the current topic and the course content. For example, “Ah yes the beach boys. I should have gotten that one”.

Blank messages sent in error will also get this label, as will truncated messages that have no relevant content (for example, “Hi [NAME]”).

Where identical duplicate messages are sent in error, an argument could be made to treat all except the first copy as Off-task. However, we have decided to follow the same principle described above for Affirmation messages (Section 2) and label each message based primarily on its own content. Thus, we give all duplicate messages the same label, based on the message content.
4.2 Interactive mode (I)

*Interactive* mode applies to messages that display explanation or reasoning about the current topic in response to an earlier message, and build on what was said in that message. The earlier message must contain some substantive content (Section 3). Explanation and reasoning are defined as for *Constructive Reasoning*, below. For a message to be labelled as *Interactive*, two additional requirements must be met: the message must be a direct response to the content of another message earlier in the discussion (not just to the linked presentation); and the message must continue the discussion from the earlier message (explicitly or implicitly). For example, “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”.

4.3 Constructive Reasoning mode (C1)

*Constructive Reasoning* mode applies to messages that display explanation or reasoning about the current topic (not simply paraphrasing or repeating). Note that reasoning does not need to be valid and explanations do not need to be correct! Explanation and reasoning includes proposing an explanation, or a cause and effect relationship; comparing or distinguishing between two or more conditions; elaborating on a point made in the linked student presentation or in a message earlier in the discussion; making a statement about the current topic and justifying it with evidence or personal examples; making a statement or asking a question about the current topic, giving reasons why the commenter thinks this way. For example, “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.”

Note that simply asking a question about whether A or B is the case does not count as reasoning. The commenter must also explain why they expect A to hold but not B, or why they expect behaviour under conditions P and Q to be the same (or different). Likewise, stating an opinion without justification is not enough; for example, “My opinion is that legacy methodologies like X and Y simply do not work. In fact I have abandoned any ideas that they could work.”

In the same way, mentioning personal experience to affirm a previous contribution does not meet this criterion if there is no reasoning involved. For example, “My experience matches the ideas presented by the authors - I’ve been involved in projects with too much budget and schedules - they were not completed well or ultimately failed.”

4.4 Constructive Extending mode (C2)

*Constructive Extending* mode applies to messages that introduce new content about the current topic to the discussion (not simply paraphrasing or repeating). New content can include a link to a video presentation; information from the presented paper that was not previously mentioned; a reference to other documents or resources about the current topic not previously mentioned; or a question, answer, or comment that is related to the current
topic, going beyond what has already been said in earlier messages, but not including any explanation or reasoning. For example, “Traditionally requirements have been difficult to nail down upfront. Why would obtaining usability requirements have more success than obtaining business requirements?”

When material from the presented paper is introduced into the discussion, annotators must use their judgement to determine whether is likely to be a paraphrase of the original presentation (and should thus be classified as Active not Constructive). References to other documents or resources must be specific enough that another discussion participant could reasonably be expected to identify the source. It is not enough to say, “I did a quick search on the Web for testing best practises and found a lot of information.” In contrast, the following would be acceptable as a reference: “[AUTHORS] referenced in Unit 3 Section 3 criticized Z along more practical lines.”

Finally, a novel question is not enough if no new content is added to the discussion. For example, “Did the author mention anything about using A and B to improve performance at the middle layer level?” However, a question that is introduced with context that contributes to the discussion is acceptable; for example, “I noticed authors of the paper did not mention much about X. What are your views on this hybrid of Y and Z?”

4.5 Affirmation messages (F)

Affirmation messages express agreement or thanks in response to an earlier message; for example, “Yes it is possible to extend the concept of X to data warehousing/ business intelligence It is a good idea for research; I will have to look into it. If we keep in mind Y. It is an interesting idea Thanks for the comments appreciated”. The earlier message must contain some substantive content (Section 3).

It is not appropriate to use this label if the affirmation relates to the presentation itself rather than the content of a discussion message. For example, “Excellent presentation; very well placed! Very clear and informative.”

4.6 Active Targeted mode (A1)

Active Targeted mode applies to messages that reference specific previous content (quoting, paraphrasing, repeating, linking, or questioning). This includes asking or answering clarification questions about specific points in the linked presentation or in earlier messages without adding any new content, explanation, or insight; paraphrasing or repeating something from the linked presentation or from earlier messages; or making connections between resources already mentioned. For example, “The presentation mentions X. Could you explain this a bit more?” and “No unfortunately there did not seem to be any physical implementation. This was a study only comparing their model to the simulation.”

The reference to previous content must be specific, unlike this message: “Thanks for your presentation. It was a very interesting and informative presentation. I would like to know how to start implementing such a technology into already existing systems. Has the authors mentioned anything like this?”
4.7 Active General mode (A2)

Active General mode applies to messages that show signs that the user is engaged with course content (explicitly or implicitly). Signs of engagement can include asking general questions (“is it useful?”); making non-specific references to earlier messages (“as others said”), the linked presentation (“great presentation”), or other unspecified documents (“I read something that said”); continuing a previous on-topic conversation (“you’re welcome”); reporting a technical issue (“the font was too fuzzy”, “background noise”); and talking about administrative matters. For example, “The presentation was well organized and covered the requirements for [this course].

If a specific named concept technology from the presentation is mentioned, the message should instead be tagged as Active Targeted. For example, “Interesting topic. Do you know of any other system which tags video with which to compare X?”.

4.8 Passive mode (P)

Passive mode relates to passive engagement with course content, such as reading messages or watching the video without responding. We do not use the Passive label because our data set does not include the relevant tracking information.

References


Appendix B  Supplementary material for Chapter 3

This appendix contains supplementary material from the peer-reviewed journal paper in Chapter 3 (Farrow et al., 2022). Figure B.1 replaces Figure 4 of Farrow et al. (2022), reproduced on p. 62, which had an error in its labelling.

Figure B.1: Update of Figure 4 from Farrow et al. (2022), now with the correct label order on the X-axis.
Table 20. Redacted Sample Messages from the Data Set with Each Framework Label

<table>
<thead>
<tr>
<th>Label</th>
<th>Sample message (redacted)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CoI phases of cognitive presence</strong></td>
<td></td>
</tr>
<tr>
<td>Triggering event</td>
<td>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]</td>
</tr>
<tr>
<td>Exploration</td>
<td>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 &quot;[Y]&quot; rather most folks in the field are &quot;bundled&quot; 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 &quot;cut&quot; of the industry roles which would serve to skew results. Sounds like you and I are very skeptical of the results. I cannot blame you. Regards [NAME]</td>
</tr>
<tr>
<td>Integration</td>
<td>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]</td>
</tr>
<tr>
<td>Resolution</td>
<td>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 person's 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</td>
</tr>
<tr>
<td>Other</td>
<td>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]</td>
</tr>
</tbody>
</table>

**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] |
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.
Appendix B. Supplementary material for Chapter 3

(d) SD of word length in letters: Coh-Metrix label DESWL1td

(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
(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 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 semicolons: LIWC label SemiC

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

(e) Number of expressions of assent: LIWC label assent
Appendix B. Supplementary material for Chapter 3

(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 poseso
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.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)
Appendix B. Supplementary material for Chapter 3

148

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)
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.
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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Features used</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coh-Metrix</td>
<td>106</td>
<td>106</td>
<td>106</td>
<td>48</td>
<td>38</td>
</tr>
<tr>
<td>LIWC</td>
<td>91</td>
<td>91</td>
<td>91</td>
<td>24</td>
<td>64</td>
</tr>
<tr>
<td>Message depth</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>First/last message</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>Number of replies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>direct + indirect</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>previous/next messages</td>
<td>–</td>
<td>–</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Count of named entities</td>
<td>–</td>
<td>relevant</td>
<td>relevant</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>Internal coherence</td>
<td>–</td>
<td>LSA</td>
<td>LSA</td>
<td>embeddings</td>
<td>LSA</td>
</tr>
<tr>
<td><strong>Metrics reported</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohen’s $\kappa$</td>
<td>0.358</td>
<td>0.63</td>
<td>0.38</td>
<td>0.72</td>
<td>0.53</td>
</tr>
<tr>
<td>Macro-averaged $F_1$</td>
<td>0.515</td>
<td>–</td>
<td>0.54</td>
<td>0.63</td>
<td>–</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cm.DESWC</td>
<td>Number of words</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>message.is.first</td>
<td>First message</td>
<td>–</td>
<td>3</td>
<td>–</td>
<td>×</td>
</tr>
<tr>
<td>3</td>
<td>liwc.posemo</td>
<td>Number of +ve emotion words</td>
<td>–</td>
<td>9</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>4</td>
<td>cm.WRDiHa</td>
<td>Meaningfulness</td>
<td>–</td>
<td>12</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>5</td>
<td>message.depth</td>
<td>Message depth in discussion</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>cm.LDTERa</td>
<td>Lexical diversity, all words</td>
<td>3</td>
<td>6</td>
<td>16</td>
<td>–</td>
</tr>
<tr>
<td>7</td>
<td>liwc.SemiC</td>
<td>Number of semicolons</td>
<td>–</td>
<td>5</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>8</td>
<td>liwc.OMark</td>
<td>Number of question marks</td>
<td>7</td>
<td>10</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>cm.WRDHYPn</td>
<td>Hypernyms for nouns</td>
<td>–</td>
<td>8</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>10</td>
<td>message.replies.direct</td>
<td>Number of direct replies</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>11</td>
<td>liwc.affect</td>
<td>Number of affective process words</td>
<td>–</td>
<td>15</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>12</td>
<td>liwc.discrep</td>
<td>Number of discrepancy words</td>
<td>–</td>
<td>7</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>13</td>
<td>liwc.money</td>
<td>Number of money words</td>
<td>10</td>
<td>11</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>14</td>
<td>message.thread.size</td>
<td>Discussion size</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>15</td>
<td>message.replies.all</td>
<td>Total number of replies</td>
<td>13</td>
<td>16</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>16</td>
<td>cm.LSASSpd</td>
<td>SD of LSA overlap in paragraph</td>
<td>–</td>
<td>19</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>17</td>
<td>cm.DESWLtd</td>
<td>SD of word length in letters</td>
<td>–</td>
<td>13</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>18</td>
<td>cm.LDVFCD</td>
<td>Lexical diversity, VOCD</td>
<td>9</td>
<td>18</td>
<td>14</td>
<td>–</td>
</tr>
<tr>
<td>19</td>
<td>cm.LDTERc</td>
<td>Lexical diversity, content words</td>
<td>5</td>
<td>–</td>
<td>15</td>
<td>–</td>
</tr>
<tr>
<td>20</td>
<td>liwc.hear</td>
<td>Number of hearing-related words</td>
<td>–</td>
<td>–</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

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 and colleagues (2019) were not previously published.
Bibliography


Farrow, E., Moore, J., & Gašević, D. (2021a). A network analytic approach to integrating multiple quality measures for asynchronous online discussions. LAK21: 11th


Supported Collaborative Learning (pp. 569–584). Springer International Publishing. https://doi.org/10.1007/978-3-030-65291-3_31. (Cited on pp. 9, 13–14, 18, 27, 129)


Ryan, R. S. (2013). The effect of online discussion forums on student engagement and perception of learning in a science course at the community college level (Doctoral dissertation). The University of Southern Mississippi. https://aquila.usm.edu/cgi/viewcontent.cgi?article=1224&context=dissertations. (Cited on p. 1)


