

UNIVERSITY OF SOUTHAMPTON  
FACULTY OF ENGINEERING AND PHYSICAL SCIENCES  
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE  
WEB AND INTERNET SCIENCE RESEARCH GROUP

**A Platform-Agnostic Model and Analysis of  
Learner Engagement within Peer-Supported  
Digital Environments: FutureLearn MOOCs  
and PeerWise**

*A thesis for the degree of  
Doctor of Philosophy in Computer Science  
(Mayflower)*

by

*Adriana Gabriela Wilde*  
ORCID 0000-0002-1684-1539

Supervisor:  
*Dr David Millard*

Examiners:  
*Dr Rebecca Ferguson* (IOT, Open University, UK)  
*Dr Mark Weal* (ECS, University of Southampton)

June 2021



*To my family, also known as the A team:  
Alexandra, my vector of unconditionally uplifting love,  
Andy, my constant in a two-decade long dream.  
This achievement is dedicated to you both!*

*(... y a mi pequeño sobrino a quien anhelo conocer: Abraham).*





UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

School of Electronics and Computer Science

Doctor of Philosophy

A PLATFORM-AGNOSTIC MODEL AND ANALYSIS OF LEARNER ENGAGEMENT  
WITHIN PEER-SUPPORTED DIGITAL ENVIRONMENTS: FUTURELEARN MOOCS AND  
PEERWISE

by **Adriana Gabriela Wilde**

Digital technologies have accelerated a conceptual shift in education from traditional face-to-face instruction towards an increasingly asynchronous, online, learner-centred paradigm. Under this paradigm, learners interact both with peers and content matter, leaving traces that can be used to characterise their learning engagement. This is the focus of a growing interest in learning analytics, particularly with data mining algorithms, of which clustering are an important class. These algorithms are however usually applied to datasets from a single platform, leading to platform-specific findings.

This thesis presents a new model of learner engagement within peer-supported digital environments that describes interactions independently of their platform, and can help make meaningful comparisons across contexts. The model was validated by applying a machine-learning approach to datasets from courses in face-to-face instruction and online. Data processed were from a total of 271,851 learners from nineteen courses from the University of Southampton between 2014-2019 on topics on archaeology, language teaching and human-computer interaction. Seventeen of these were massive open online courses (MOOCs), and the remaining two were in a face-to-face setting that included the use of PeerWise as a peer-supported digital environment.

Feature engineering was performed on timestamped digital traces of activity using this new model, producing sixteen feature files with up to 78 features per learner, which were subjected to the clustering algorithms Expectation Maximization, Simple k-Means and X-Means with k values varying from two to ten. Highly-interpretable clusters were identified by X-Means on dialogic features from datasets from both platforms, allowing for a meaningful comparison of learner engagement across environments. In particular, engagement in both platforms was found to fall in four main activity classes ranging from asocial to fully active social learners; although nuanced behaviours were also evidenced. Learning design was found to affect the composition of these clusters, and when free of behavioural constraints, learners in the face-to-face environment evidenced the same types of behaviours as those online.

**Keywords:** PeerWise, MOOC, learning analytics, learner engagement, clustering.



# Table of Contents

<b>Abstract</b>	<b>v</b>
<b>Table of Contents</b>	<b>vii</b>
<b>List of Tables</b>	<b>xi</b>
<b>List of Figures</b>	<b>xiii</b>
<b>Declaration of Authorship</b>	<b>xix</b>
<b>Acknowledgements</b>	<b>xxiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation for this thesis . . . . .	3
1.2 Purpose of this research . . . . .	5
1.3 Thesis outline . . . . .	5
<b>2 Literature Review</b>	<b>9</b>
2.1 Learners and engagement . . . . .	10
2.1.1 Behaviour change for engagement . . . . .	12
2.1.2 Experimenting in educational research . . . . .	13
2.1.3 e-Learning and Learning at Scale . . . . .	15
2.1.4 Peer-learning: Experiential and conversational learning . . . . .	16
2.2 Massive Open Online Courses . . . . .	18
2.2.1 FutureLearn . . . . .	27
2.2.1.1 Architecture . . . . .	28
2.2.1.2 Design . . . . .	29
2.3 PeerWise . . . . .	33
2.3.1 Multiple-Choice Questions . . . . .	35
2.3.2 PeerWise affordances . . . . .	36
2.3.3 Gamification elements in PeerWise . . . . .	36
2.4 Other peer-supported digital environments for face-to-face instruction . . . . .	38
2.5 Learning analytics, educational data mining and academic analytics . . . . .	40
2.5.1 Feature Engineering . . . . .	42
2.5.2 Clustering . . . . .	44
2.6 Interval Algebra . . . . .	44
2.7 Information retrieval terminology . . . . .	46

2.7.1	Multi-Class Approach Considerations . . . . .	46
2.8	Summary . . . . .	47
<b>3</b>	<b>Methodological framework</b>	<b>51</b>
3.1	High-level description of the approach . . . . .	51
3.2	A quantitative approach for data science . . . . .	53
3.2.1	Ask question . . . . .	53
3.2.2	Collect data . . . . .	54
3.2.3	Clean data . . . . .	59
3.2.4	Define new features . . . . .	60
3.2.5	Deploy . . . . .	62
3.3	Summary . . . . .	62
<b>4</b>	<b>A platform-agnostic model of learner interactions</b>	<b>65</b>
4.1	e-tivities . . . . .	67
4.2	Types of <i>e-tivities</i> . . . . .	69
4.3	Communicative e-tivities . . . . .	73
4.3.1	A practical scenario . . . . .	75
4.3.2	Limitations of the chat representation . . . . .	79
4.3.3	Communicative e-tivities as n-order replies . . . . .	81
4.4	Non-communicative e-tivities . . . . .	85
4.4.1	Time beyond a timestamp: Intervals . . . . .	86
4.4.2	A practical scenario . . . . .	86
4.4.3	Putting it all together . . . . .	89
4.5	Limitations of the model . . . . .	91
4.6	Summary of the model . . . . .	94
4.7	Conclusion of this chapter . . . . .	97
<b>5</b>	<b>Peer-learning online within FutureLearn MOOCs</b>	<b>99</b>
5.1	Motivation and context . . . . .	100
5.2	Datasets . . . . .	101
5.2.1	Learning design changes . . . . .	102
5.3	A heuristic approach to discussion analytics . . . . .	104
5.3.1	Applying the heuristic by Chua et al. (2017) . . . . .	107
5.4	Feature engineering on FutureLearn MOOC data . . . . .	109
5.4.1	Dialogic features . . . . .	110
5.4.2	Interval features . . . . .	110
5.4.3	Badge features . . . . .	112
5.4.4	Other features . . . . .	112
5.5	Selecting a feature set . . . . .	114
5.5.1	Why not Principal Component Analysis? . . . . .	114
5.5.2	Semantically-chosen features . . . . .	115
5.6	Clustering algorithm on MOOC features . . . . .	116
5.6.1	Which clustering algorithm? . . . . .	118

5.6.2	How many clusters? . . . . .	119
5.7	Results . . . . .	121
5.7.1	Size and coherence of resulting clusters . . . . .	121
5.7.2	Semantically chosen names for clusters in both MOOCs . . . . .	123
5.7.3	Semantically chosen names for clusters in each run of a MOOC . . . . .	128
5.7.4	Distribution of learners across the newly-named clusters . . . . .	131
5.7.5	Comparing against the learner types as per Chua's heuristic . . . . .	134
5.8	Summary and conclusion for this chapter . . . . .	138
<b>6</b>	<b>Peer-learning in face-to-face instruction mediated by PeerWise</b>	<b>141</b>
6.1	Motivation and context . . . . .	142
6.2	Datasets and learning design . . . . .	144
6.2.1	The first cohort: class of 2015/16 . . . . .	145
6.2.2	Learning design changes for the second cohort: class of 2016/17 . . . . .	147
6.3	Modelling interactions in PeerWise . . . . .	151
6.4	Features on PeerWise data . . . . .	153
6.5	Clustering on PeerWise features . . . . .	157
6.5.1	Size and coherence of resulting clusters . . . . .	157
6.5.2	Semantically chosen names for clusters in both courses . . . . .	159
6.6	Reflecting back to MOOC analysis . . . . .	165
6.7	Summary and conclusion for this chapter . . . . .	168
<b>7</b>	<b>Conclusion</b>	<b>171</b>
7.1	Summary of this thesis . . . . .	172
7.2	Answering the research questions . . . . .	173
7.3	Contributions of this thesis . . . . .	178
7.3.1	A platform-agnostic model for analysis of learner engagement . . . . .	178
7.3.2	Anonymised datasets with up to 78 features on sixteen MOOCs and 72 features on two face-to-face courses . . . . .	180
7.3.3	Profiles of learner engagement in MOOCs and PeerWise . . . . .	180
7.3.4	A critique of Chua, Tagg, Sharples, and Rienties (2017) . . . . .	180
7.3.5	A theory of behavioural constraint . . . . .	181
7.3.6	Publications and talks . . . . .	181
7.4	Limitations . . . . .	183
7.5	Future work . . . . .	184
7.5.1	Interval features . . . . .	184
7.5.2	Principal Component Analyses . . . . .	185
7.5.3	PeerWise adoption when not aligned with assessment . . . . .	186
7.6	Concluding remarks . . . . .	186
	<b>References</b>	<b>189</b>
<b>A</b>	<b>Ethics and Research Governance Online 2 at Southampton</b>	<b>A-1</b>
<b>B</b>	<b>Data Protection Impact Assessment</b>	<b>B-1</b>

<b>C</b>	<b>Data Management Plan</b>	<b>C-1</b>
<b>D</b>	<b>Additional details for MOOCs datasets</b>	<b>D-1</b>
<b>E</b>	<b>Synthetic data for the example in Figure 6.1 under PeerWise</b>	<b>E-1</b>
<b>F</b>	<b>Coursework specification for participation in PeerWise (COMP2213)</b>	<b>F-1</b>
<b>G</b>	<b>Step-centred analysis of the first run of Understanding Language</b>	<b>G-1</b>
<b>H</b>	<b>Clustering on Interval Features in FutureLearn MOOCs</b>	<b>H-1</b>
	H.1 Expectation Maximisation clustering with interval features . . . . .	H-2
<b>I</b>	<b>Principal Component Analyses</b>	<b>I-1</b>
	I.1 PCA for Understanding Language (run 1) . . . . .	I-2
	I.2 PCA on Portus (run 6) . . . . .	I-4
	I.3 PCA on PeerWise data (course 12710, first cohort of COMP2213) . . . . .	I-6
	I.4 PCA on PeerWise data (course 14715, second cohort of COMP2213) . . . . .	I-8
<b>J</b>	<b>Detailed accuracy for classification on clusters found with X-Means</b>	<b>J-1</b>
	J.1 Results with $k = 4$ . . . . .	J-1
	J.1.1 Portus . . . . .	J-1
	J.1.2 Understanding Language . . . . .	J-2
	J.1.3 First cohort with PeerWise (12710) . . . . .	J-3
	J.1.4 Second cohort with PeerWise (14715) . . . . .	J-4
	J.2 Results with $k = 7$ . . . . .	J-5
	J.2.1 Portus . . . . .	J-5
	J.2.2 Understanding Language . . . . .	J-6
	J.2.3 First cohort with PeerWise (12710) . . . . .	J-7
	J.2.4 Second cohort with PeerWise (14715) . . . . .	J-9
<b>K</b>	<b>Clustering FutureLearn MOOCs with X-Means and <math>k=4</math></b>	<b>K-1</b>
<b>L</b>	<b>Clustering individual runs of FutureLearn MOOCs with X-Means and <math>k=7</math></b>	<b>L-1</b>
	L.1 Portus . . . . .	L-2
	L.2 Understanding Language . . . . .	L-9

# List of Tables

2.1	Table of interventions . . . . .	14
2.2	Summary of similarities and differences between c-MOOCs and x-MOOCs. Adapted from Cobos, Wilde, and Zaluska (2017) . . . . .	19
2.3	Summary of reviewed categorisations of MOOC learners in the literature . . . . .	21
2.4	Summary of the FutureLearn archetypes . . . . .	31
2.5	Summary of relevant quantitative studies of PeerWise learners in the literature . . . . .	34
2.6	Badges currently available in PeerWise for student participation. . . . .	39
2.7	Interval algebra: the thirteen possible relations (adapted from Allen (1983) and Hunsdale, Chuckravanen, Daykin, and Seam (2017)). . . . .	45
4.1	Complementary interpretations of views on learning activities . . . . .	69
5.1	Enrolled and active learners per offering of each course (run) as extracted from the datasets. (Source of starting dates per run: Class Central <sup>5</sup> ). . . . .	102
5.2	Steps per week in each run for both courses, as extracted from the datasets. . . . .	103
5.3	Absolute counts per social learner group in the first run of the Personal Finance MOOC calculated by Chua et al. (2017) . . . . .	105
5.4	Absolute counts per social learner group in the categorisation by Chua et al. (2017) . . . . .	107
5.5	Dialogic features engineered (as informed by the model in Chapter 4) . . . . .	110
5.6	Interval features engineered (as informed by the model in Chapter 4) . . . . .	111
5.7	Extract of the step-activity file associated to the toy example MOOC from Figure 4.10), built to illustrate the calculation of interval features . . . . .	111
5.8	Sequences of steps and their Event_types for each learner in the toy example from Figure 4.10 . . . . .	111
5.9	Values of the interval features for each learner in the toy example . . . . .	111
5.10	“Badge” features engineered (inspired from PeerWise badges in Chapter 6) . . . . .	112
5.11	Other features extracted for MOOC learners . . . . .	112
5.12	Semantic classes for the clusters found by X-Means in Portus (for each run) . . . . .	129
5.13	Numbers of learners in each of the semantic classes for the clusters found by X-Means in Portus . . . . .	129
5.14	Semantic classes for the clusters found by X-Means in Understanding Language (for each run) . . . . .	130
5.15	Numbers of learners in each of the semantic classes for the clusters found by X-Means in Understanding Language . . . . .	130
5.16	Summary comparative table of clusters . . . . .	138

6.1	Differences in assessment design between the two offerings of Interaction Design under study. Note that there are marks given to participation in PeerWise in the first deployment, which is absent in the second. . . . .	148
6.2	Comparative statistics and other characteristics of the consecutive offerings of the Interaction Design module. . . . .	149
6.3	Files in the 2015/16 dataset for PeerWise. All the listed files have extension .csv. . . . .	150
6.4	Files in the 2016/17 dataset for PeerWise. All the listed files have extension .csv. . . . .	150
6.5	Features extracted from the PeerWise dataset . . . . .	154
6.6	Features extracted from the Assessment data dataset (including the Wiki, where groups allocation was published) . . . . .	154
6.7	Features engineered from the PeerWise dataset for a given student $s$ (with a unique $User\_ID$ as per Table 6.5). . . . .	156
6.8	Comparison of numbers of learners in each of the semantic classes for the clusters found by X-Means in both peer-supported digital environments . .	166
6.9	Percentages of learners in each of the semantic classes for the clusters found by X-Means in both peer-supported digital environments. Due to rounding, percentages may not add up to 100%. . . . .	166
6.10	Comparative table of clusters found in PeerWise data with X-Means ( $k = 7$ ) in each cohort of Interaction Design (PeerWise courses 12710 and 14715).169	
7.1	Summary comparative table of clusters . . . . .	175
7.2	A summary table of comparisons between clusters found amongst Interaction Design students (more details in Table 6.10). . . . .	176
7.3	Percentages of learners in each of the semantic classes for the clusters found by X-Means in both peer-supported digital environments. Due to rounding, percentages may not add up to 100%. . . . .	178
D.1	Files in the Portus MOOC dataset . . . . .	D-2
D.2	Files in the understanding-language MOOC dataset (runs 1..8) . . . . .	D-3
D.3	Files in the understanding-language MOOC dataset (runs 9..11) . . . . .	D-4
D.4	Summary table of entries per file in the Portus MOOC dataset per run . . .	D-4
D.5	Summary table of entries per file in the Understanding Language MOOC dataset per run . . . . .	D-5
D.6	Dates in the portus MOOC dataset . . . . .	D-6
D.7	Dates in the understanding-language MOOC dataset . . . . .	D-6
E.1	Files in the test dataset for PeerWise . . . . .	E-1
E.2	Contents of <code>Users_test.csv</code> for the example in Figure 6.1. . . . .	E-2
E.3	Contents of <code>Questions_test.csv</code> for the example in Figure 6.1. (some fields omitted) . . . . .	E-2
E.4	Contents of <code>Comments_test.csv</code> for the example in Figure 6.2. . . . .	E-2
E.5	Contents of <code>Replies_test.csv</code> for the example in Figure 6.1. . . . .	E-3
E.6	Contents of <code>Ratings_test.csv</code> for the example in Figure 6.2. . . . .	E-3
E.7	Contents of <code>Answers_test.csv</code> for the example in Figure 6.2. . . . .	E-3



# List of Figures

1.1	Computer Science Education: the main research theme in my research . . .	3
1.2	Research Framework Diagram . . . . .	6
1.3	Thesis organisation . . . . .	7
2.1	Multi-level categorisation model of conceptions of teaching . . . . .	17
2.2	Growth of MOOCs since 2012 (as reported on Class Central) . . . . .	18
2.3	Archetype survey question in a FutureLearn MOOC . . . . .	30
2.4	PeerWise leaderboard example (from <a href="https://peerwise.cs.auckland.ac.nz/docs/students/">https://peerwise.cs.auckland.ac.nz/docs/students/</a> ). . . . .	37
2.5	PeerWise badges (from <a href="https://peerwise.cs.auckland.ac.nz/docs/students/">https://peerwise.cs.auckland.ac.nz/docs/students/</a> ). . . . .	37
2.6	Learning Analytics is a multidisciplinary field . . . . .	41
3.1	High-level view of scenarios spanning two different educational contexts .	52
3.2	Data science pipeline applied in this study . . . . .	53
3.3	Associated Files for the FutureLearn MOOCs datasets (part I) . . . . .	56
3.4	Associated Files for the FutureLearn MOOCs datasets (part II). Generated by FL only for run 6 of Portus and runs 8 and 9 of Understanding Language)	57
3.5	Associated Files for the PeerWise datasets (I) . . . . .	58
3.6	Associated Files for the PeerWise datasets (II). Some fields are omitted for clarity . . . . .	59
4.1	Basis of the platform-agnostic model of learner interactions . . . . .	66
4.2	Silhouette of a partly-occluded dancer: a metaphor . . . . .	67
4.3	Learner engagement within a platform: conversing, producing and consuming . . . . .	70
4.4	Examples of learning activities in digital environment according to the dimensions peer vs content and level of activity . . . . .	72
4.5	Hypothetical scenario with sets of conversations amongst learners . . . . .	77
4.6	Alternative representation of the illustrated scenario, showing the contextual relationships between learners and their posts, comments and replies. To the left of each post, a code is given to indicate their types of communicative e-tivity: starting post (SP), lone post (LP), first reply (FR), additional reply (AR) and initiator's reply (IR). . . . .	78

4.7	Timeline of posts, comments and replies in the scenario introduced in Figure 4.5, showing various types of interactions, and the times when they took place. *Note that post in $t_{12}$ is still a <b>lone post</b> , since even though it has sparked a comment and a reply, these are from the initiator, $l_3$ , Cam. . . . .	80
4.8	Timeline of e-tivities in the illustrated scenario, showing various types of interactions. Each communicative e-tivity falls into one of five categories: starting posts, lone posts, first replies, additional replies, and initiator's replies. *Note that post $\langle \langle p_7, m_{\perp} \rangle, l_3, t_{12} \rangle$ is still a <b>lone post</b> , since its only comment $\langle \langle c_5, p_7 \rangle, l_3, t_{13} \rangle$ and reply $\langle \langle r_5, c_4 \rangle, l_3, t_{16} \rangle$ are both from the initiator, $l_3$ . . . . .	83
4.9	Alternative representation of the illustrated scenario, showing contextual relationship between posts, comments and replies. Here, a learner $l_i$ makes a post $p_j$ (at depth=1 in the tree), which in turn may raise a comment $c_k$ from learner $l_x$ (at depth=2). Learner $l_y$ then makes a reply $r_m$ to a comment (at depth=3). . . . .	84
4.10	Timeline of non-communicative e-tivities in a second scenario, showing various types of behaviour amongst learners $l_1, l_2, l_3$ working in activities $w_1, w_2, w_3$ over time. Starts and ends of activity are shown by empty and full circles. The absence of a full circle indicates an ongoing or abandoned activity (so a finishing time of $t_{\infty}$ is assigned.) . . . . .	87
4.11	Forest representation of the communicative activities in the second scenario, showing contextual relationships between zero- first- and second-order replies. Here, learner $l_1$ makes posts $p_1$ and $p_2$ (at depth=1), followed by a comment $c_1$ on $p_1$ from $l_2$ (at depth=2), which in turn is replied by the initiator, $l_1$ (at depth=3) with reply $r_1$ which is then replied to by learner $l_3$ with $r_2$ . . . . .	89
4.12	Timeline of the communicative activities of the scenario presented in Figure 4.11. . . . .	90
4.13	Timeline of communicative and non-communicative e-tivities in the second scenario, involving learners $l_1, l_2, l_3$ . Circles indicate the starts and ends of a non-communicative activity. A square indicates a communicative activity. . . . .	91
4.14	Timeline of the non-communicative e-tivities shown in Figure 4.10, interspersed with the communicative activities of the scenario. . . . .	92
4.15	Screenshot in the Whatsapp messaging app whilst someone is typing . . .	93
5.1	Distribution of social learners in the Personal Finance MOOC (from Table 5.4), in descendent order by number of learners in each category. Note that the largest category is the one comprising active social learners.	106
5.2	Portus runs according to the heuristic (from Table 5.4). . . . .	108
5.3	Understanding Language runs according to the heuristic (from Table 5.4)	108
5.4	Dialogic features for Understanding Language (all runs) and Portus (all runs) with k-Means . . . . .	117
5.5	Interval features for Understanding Language (all runs) and Portus (all runs) with k-Means . . . . .	117

5.6	Interval features for Understanding Language (all runs) and Portus (all runs) with k-Means . . . . .	118
5.7	Dialogic features for Portus (all runs) with several clustering algorithms . . . . .	120
5.8	Dialogic features for Understanding Language (all runs) with several clustering algorithms . . . . .	120
5.9	Confusion matrix plots for the clusters found by the X-Means clustering algorithm on the Portus MOOC (all runs combined), with $k = 7$ . . . . .	122
5.10	Confusion matrix plots for the clusters found by the X-Means clustering algorithm on the Understanding Language MOOC (all runs combined), with $k=7$ . . . . .	122
5.11	Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the Portus MOOC (all runs), with $k = 7$ . . . . .	125
5.12	Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on Understanding Language MOOC (all runs), with $k = 7$ . . . . .	127
5.13	Distribution of social learners across on Portus according to the clusters found by X-Means ( $k = 7$ ) and interpreted using the classification by Chua et al. (2017). . . . .	132
5.14	Distribution of social learners across on Understanding Language according to the clusters found by X-Means ( $k = 7$ ) and interpreted using the classification by Chua et al. (2017). . . . .	133
5.15	Distribution of social learners across on Portus according to the clusters found by X-Means ( $k=7$ ) and aggregated guided by Chua's classification . . . . .	135
5.16	Distribution of social learners across on Understanding Language according to the clusters found by X-Means ( $k=7$ ) and aggregated guided by Chua's classification . . . . .	135
5.17	Confusion matrix plot for the clusters found by the X-Means clustering algorithm on the Portus MOOC (all runs), with $k = 7$ , against those categories identified by applying Chua's heuristic . . . . .	137
6.1	A simplified test case with three students in PeerWise, creating questions, comments and replies. In this example, student $s_i$ makes question $q_{i,j}$ , which in turn raises comment $c_{i,j,k}^t$ by learner $s_t$ , and student $s_p$ gives a reply $r_{i,j,k,l}^p$ to the comment. . . . .	151
6.2	Test case graph of students, questions, comments and replies as per Tables E.4 and E.5. The supra-index notation for comments and replies allows to keep track of comments' and replies' authors. As an example, the additional lines show that student $s_2$ authored comment $c_{1,1,1}^2$ (the solid blue line) and reply $r_{1,1,2}^2$ (the red dashed line) . . . . .	152
6.3	Confusion matrix plots for the clusters found by the X-Means clustering algorithm on the first cohort data, with $k = 7$ . . . . .	158
6.4	Confusion matrix plots for the clusters found by the X-Means clustering algorithm on the second cohort data, with $k=7$ . . . . .	158
6.5	Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the first cohort (all runs), with $k = 7$ . . . . .	161

6.6	Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the second cohort (14715), with $k = 7$ . . . . .	163
6.7	Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on PeerWise course 12710, with $k=4$ . . . . .	164
6.8	Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on PeerWise course 14715, with $k=4$ . . . . .	165
7.1	Research Framework Diagram, revisited . . . . .	179
G.1	WEKA Explorer visualisation of features extracted from the step-activity file associated to the first run of the Understanding Language MOOC . . .	G-2
G.2	Distinct count of completion time for each step, organised by step type . .	G-4
G.3	Visited steps in the Understanding Language MOOC (run 1), per step type	G-5
G.4	Completed steps in the Understanding Language MOOC (run 1), per step type . . . . .	G-5
K.1	Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the Portus MOOC (all runs), with $k=4$ . . . . .	K-2
K.2	Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the Understanding Language MOOC (all runs), with $k=4$ . .	K-2
L.1	Confusion matrix plots for the clusters found by the X-Means clustering algorithm on Portus MOOC (all runs), with $k = 7$ . . . . .	L-2
L.2	Portus 1 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the first run of Portus, with $k=7$ . . . . .	L-3
L.3	Portus 2 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the second run of the Portus MOOC, with $k=7$ . . . . .	L-4
L.4	Portus 3 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the third run of the Portus MOOC, with $k=7$ . . . . .	L-5
L.5	Portus 4 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the fourth run of the Portus MOOC, with $k=7$ . . . . .	L-6
L.6	Portus 5 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the fifth run of the Portus MOOC, with $k=7$ . . . . .	L-7
L.7	Portus 6 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the sixth run of the Portus MOOC, with $k=7$ . . . . .	L-8
L.8	Confusion matrix plots for the clusters found by the X-Means clustering algorithm on the Understanding Language MOOC (for each run), with $k = 7$ . . . . .	L-9

---

L.9	Understanding Language 1 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the first run of the Understanding Language MOOC, with k=7 . . . . .	L-10
L.10	Understanding Language 2 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the second run of the Understanding Language MOOC, with k=7 . . . . .	L-11
L.11	Understanding Language 4 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the fourth run of the Understanding Language MOOC, with k=7 . . . . .	L-12
L.12	Understanding Language 5 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the fifth run of the Understanding Language MOOC, with k=7 . . . . .	L-13
L.13	Understanding Language 6 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the sixth run of the Understanding Language MOOC, with k=7 . . . . .	L-14
L.14	Understanding Language 7 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the seventh run of the Understanding Language MOOC, with k=7 . . . . .	L-15
L.15	Understanding Language 8 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the eighth run of the Understanding Language MOOC, with k=7 . . . . .	L-16
L.16	Understanding Language 9 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the ninth run of the Understanding Language MOOC, with k=7 . . . . .	L-17
L.17	Understanding Language 10 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the tenth run of the Understanding Language MOOC, with k=7 . . . . .	L-18
L.18	Understanding Language 11 clusters' semantics: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the on the eleventh run of the Understanding Language MOOC, with k=7 . . . . .	L-19



# Declaration of Authorship

I, **Adriana Gabriela Wilde**, declare that the thesis entitled

**A Platform-Agnostic Model and Analysis of Learner Engagement within  
Peer-Supported Digital Environments: FutureLearn MOOCs and PeerWise**

and the work presented in it are my own and have been generated by me as the result of my own original research. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published, as seen in the list of selected publications.

**Signed:** Adriana Wilde

**Date:** June 2021

## List of Selected Publications

Peer-reviewed publications that contain work described as part of this thesis are listed below. Other work is listed at: <http://scholar.google.co.uk/citations?user=-xi8>.

### Journal Papers

1. S. Snow, A. Wilde, m.c. schraefel, and P. Denny (2019) “**A discursive question: Supporting student-authored multiple-choice questions through peer-learning software in non-STEMM disciplines**”. *British Journal of Educational Technology*, 50 (4), pp. 1815–1830.

### Peer-Reviewed International Conferences (Posters)

2. A. Wilde and D. Millard (2020) “**Choosing between wider participation or quality of interactions: A study of learner engagement within PeerWise**”. In *The United Kingdom and Ireland Computing Education Research (UKICER) conference from the UK ACM Special Interest Group in Computing Science Education*. Organised by the University of Glasgow, online, 3-4 September 2020.

### Peer-Reviewed Contributions to International Workshops

3. R. Cobos, A. Wilde and E. Zaluska (2017) “**Predicting attrition from Massive Open Online Courses in FutureLearn and edX**”. In *FutureLearn data: what we currently have, what we are learning and how it is demonstrating learning in MOOCs*. Workshop at the 7th International Learning Analytics and Knowledge Conference, 13–17 March, Simon Fraser University, Vancouver, Canada. pp. 1–20.
4. A. Wilde (2016) “**Understanding persuasive technologies to improve completion rates in MOOCs**”. In the *HCI and the Educational Technology Revolution – an HCI Educators workshop* at the 2016 International Conference on Advanced Visual Interfaces (AVI 2016), 7 June, Bari, Italy.
5. A. Wilde, E. Zaluska and D. Millard (2015) “**Student success on face-to-face instruction and MOOCs: What can learning analytics uncover?**” In the *Web Science Education Workshop* at the ACM Web Science Conference, 28 June, Oxford, UK.



## Peer-Reviewed Contributions to National Events

6. **A. Wilde** and S. Snow (2018) “**Addressing challenges in assessing Human-Computer Interaction at scale**”. In *Computing Education Practice conference*, 11 – 12 January, Durham, UK.
7. S. Snow and **A. Wilde** (2017) “**Supporting authoring of multiple-choice questions in human-computer interaction using PeerWise**”. In *What Works in Assessment and Feedback: Simply Better conference*, 14 September, Southampton, UK.



# Acknowledgements

Firstly I would like to thank my School for selecting me as a Mayflower Scholar, as otherwise I would not have been able to even start my PhD, let alone get to where I am today. In addition to the funding to pursue my research, I was given the opportunity to teach large classes at a top department, alongside experienced academics from whom I learned so much, notably: Ed Zaluska, Tim Chown, Lie-Liang Yang, Jeff Reeve (RIP), Kirk Martinez, Corina Cirstea, Jon Hare, Bob Walters, Gary Wills, Denis Nicole, Federica Paci and several others. It has been an honour to have you as colleagues. Special thanks to Paul Lewin, for keeping an eye on me over the years as I matured as an educator, a researcher, and a whole person. What you taught me helped me navigate old challenges and embrace new opportunities. For your time, advice and support, thanks!

I thank my supervisors: firstly, Ed Zaluska for taking a leap of faith with me, setting me off in my PhD journey and letting me become an independent scholar. For this, I will be forever grateful. Thanks to Dave Millard for teaching me to “kill my darlings” and not to have “two clocks”. For your guidance was instrumental for me to refocus my research so that these pages before you no longer look like a mere collection of explorations but a *Thesis*, with a coherent narrative about my contributions and how I got there. Immense thanks for helping me pick up the pieces, dig deeper and with intention.

Due thanks go to those who played important roles on both procedural and intellectual aspects of my PhD: my examiners Mark Weal and Rebecca Ferguson produced thoughtful and useful examination reports; Les Carr and Megan Chan welcomed me warmly back in the lab after nearly two years at St Andrews; Brian Pickering and Alison Knight advised me on my applications for Ethics Research Governance Online and Data Protection Impact assessments; Michael Whitton and Isobel Stark provided feedback on my Data Management Plan; Nick Gibbins helped me recover assessment data for Chapter 6, and gave me constructive feedback on the rigour of the model described in Chapter 4; Kate Borthwick facilitated access to the FutureLearn data I used in Chapter 5; and Gary Wills for chairing my viva voce examination. I am also indebted to Ben Sanders and Martin Broad from my new home at the University of Winchester, for all the arrangements that allowed me to take a short sabbatical to finish these corrections.

Thanks to Phil Tubman for our chats on his research, which would have not happened were it not for the FutureLearn Academic Network (FLAN). Other FLAN-goers who with I grew in understanding of this field were Shi-Min Chua, Tina Papathoma, Monty King, Lisa Harris and others. Thanks to Mike Sharples for coordinating these meetings and to Rebecca Ferguson and Eileen Scanlon for continuing this work after his retirement.

I am grateful to my past students too, especially those in my Interaction Design classes. For their engagement at the time, in PeerWise specifically, interactions which I studied and present in Chapter 6. For the help from my co-authors Steve Snow (who with I designed learning activities and assessment) and Paul Denny (for his explanations on the less widely documented inner works of PeerWise, his expert eye and keen interest in my writing). For their boundless *Down-Under* energy, particularly Paul's, including his flexibility in arranging Zoom calls despite the 13 hours' difference. *Thanks heaps!*

I also want to extend my gratitude to others with whom I wrote during my PhD (including some former students), even if my part in these collaborations may not have made it to this version of the thesis. Thanks for the perspectives each of you brought into our writing. Not only the products were better than they would have been had I been on my own, I have truly benefited from the mutual accountability of writing together. So I thank Pireh Pirzada, Nicolas Zurbuchen, David Harris-Birtill, Gayle Doherty, Pascal Bruegger, Olivia Ojuroye, Adalberto Simeone, Alan Dix, Chris Evans, Joe Maguire, Ana Vasilchenko, Marie Devlin, Manuel León-Urrutia, Su White, Miguel Ballesteros, Vesna Perisić and some others. Meeting deadlines was easier knowing that I was not alone.

In the final weeks of writing, I had virtual company with Cathy Mazak's global tribe of womxn who with I kept my momentum. Especially Rocío Caballero-Gill (the coach who "rocs"), Cláudia Soares, Lauren Braun-Strumfels, Sarah Charnes, Dianna Dempsey, Stephanie Nutting, Jocelyn Curtis-Quick, Lindsay Maurer Braun, Silvie Huijben, Margaret Foster, Donita Brown, Ami Stearns, Chiara Cecalupo, Heidi Cephus and Marie Kolbenstetter for their encouragement and community wisdom. I'm glad I found you!

However I have never been alone. Throughout these years I had the unwavering support of my friends: Lizeth Avendaño, Ángeles Camacho, Cherril Norrie, Soraya Nweihed, Claudia León, Marián Barrios, Dragica Kostić, David Littlehales, Katie Hilditch, Dorota Filipczuk, Jon Hare, Massimo Mecella, Jess Spurrell, Jen Forrester, Clare Hooper, Clare Hutton, Rachel Cooper, Zillah Abbt, Iman Naja, Erick Oliveros, Elizabeth Morales, Javier Bustos and Olja Rastić. Thank you for opening the doors of your homes and hearts to me, whether I needed a place to stay or a shoulder to cry on, a hot meal or a stiff drink, an hour-long phone call or a silent walk. Regardless of whether you believe in God, I see Him in your kindness and unconditional love towards me in my brokenness, because you have seen me at my worst and yet were able to bring out the best in me. I love you.

Above all though, I love my family, and I am infinitely grateful to have you. Alexandra, for your boundless creativity, sense of fun, self-drive and kindness, and Andy, for your stoic perseverance, your unrivalled wayfinding skills through Stack Overflow, and your ability to conjure up bash magic. All things I needed to learn and master, particularly towards the end of the pursuit of this goal. But also for much, much, much more (not least, all the sausage casseroles). We have been through *too* much since I started this work, including huge health scares (two), deaths of loved ones (three), and separations and tears (too many to count!) But we are together now, more than ever, so I thank you for bearing with me through this all. Together we can overcome anything!

Last but not least, thanks to my mother, Carmen Guédez, the very first doctor from her home town of Guararute, who continues to inspire me and heal me despite the seven thousand miles separating us. *iGracias por todo mamá!*

There are others who I do not mention in this already long list, either because my memory fails me or because my gratitude cannot be justly expressed in words. Either way, forgive me, and be happy that I made it this far, against all odds.



## Introduction

*“Isn’t it strange how princes and kings,  
and clowns that caper in sawdust rings,  
and common folk, like you and me,  
are builders for eternity?  
Each is given a bag of tools;  
a shapeless mass; a book of rules.  
And each must fashion, ere life is flown,  
a stumbling block, or a stepping stone.”*

Robert Lee Sharpe  
(b. 14 August 1872 – d. 19 April 1951),  
A BAG OF TOOLS, circa 1929

Though written nearly a century ago, the words in the epigraph, by the American poet Robert Lee Sharpe are still relevant today, especially in the context of education. Individuals from diverse backgrounds and interests each manifest their use ‘a *bag of tools*’ and rules together with their raw talents, to arrive perhaps at one of two possible outcomes: either ‘stumbling’ or a ‘stepping forwards’, the old-age dichotomy of attrition versus progression and success.

A more recent addition to that proverbial bag of tools, are those mediated through digital technologies. Their explosive growth and diversity has been catalysed by an increased affordability of devices with greater connectivity and computing power than ever before. Many of these tools are ripe for assisting educators at a time of radical societal change, such as those used to support interactions amongst learners, as well as those of learners with their learning resources. Further, unprecedented demands for social distancing practices across the globe over extended periods add to already high expectations for around-the-clock access to educational resources and support.

Against this background, practitioners in higher education institutions have found themselves “pivoting”, redesigning their courses, updating their methods of delivery and assessment, and adopting digital environments to replace or complement a provision which had previously followed predominantly a face-to-face instruction model. Though not all of the engagement activity of learners in face-to-face instruction is observable, some can be used as a valuable proxy for actual engagement. In particular, the digital interactions amongst learners, as well as those with their educational content, are valuable traces of the elusive true engagement. If, as I argue above, educational and societal trends result in an increase of the proportion of digital interactions, then it is imperative to study them by applying knowledge and understanding gained from research on online learning.

The main motivation behind the research activities undertaken throughout my candidature has been my interest in technology to facilitate and understand learning success in its many manifestations. This lifelong dual interest in computing and education led me to study diverse aspects of human-centered computing, from the use of computational methods to make sense of human behaviour<sup>1</sup>, people’s attitudes<sup>2</sup> to technology in general<sup>3</sup> and computing in particular<sup>4</sup>. I have also explored how technology could support positive behaviours and be persuasive<sup>5</sup>, and how the use of new technologies in learning need to still be human-centred<sup>6</sup>. Finally, I have also studied Massive Open Online Courses (MOOCs) data in more ways than those explicitly related to this thesis<sup>7</sup>. These themes are summarised in Figure 1.1.

Researching whilst teaching in higher education gave me opportunities to inform my practice with my research, and also my research with my practice. I was able to introduce some innovations in my practice which were born out of what I had been researching, and conversely, some of my research emerged from the implementations of ideas for learning activities in my practice<sup>8</sup>. For example, using clickers in my classes<sup>9</sup>,

---

<sup>1</sup>For example, in using sensor data for human activity classification (Pirzada, White, and Wilde, 2018), fall detection (Zurbuchen, Wilde, and Bruegger, 2021; Zurbuchen, Bruegger, and Wilde, 2020), and smart home technologies (Pirzada, Wilde, Doherty, and Harris-Birtill (2021); Bruegger, Wilde, and Guibert (2020); Ojuroye, Torah, Beeby, and Wilde (2017); Wilde, Ojuroye, and Torah (2015)).

<sup>2</sup>As expressed, for example, in the sentiment of their comments in massive open online courses (Wilde and Wang, 2017), explored by my student Jing Wang (2017) in her MSc project.

<sup>3</sup>I ran a bilingual survey of students’ attitudes towards smartphones in their studies (Wilde, 2015a).

<sup>4</sup>Explored through work on gender balance in computing (Wilde and Rastić-Dulborough, 2017).

<sup>5</sup>I have discussed the use of persuasive technologies for behaviour change, and how learning analytics could help tailor interventions to promote perseverance in MOOC learning (Wilde, 2016).

<sup>6</sup>Including education via virtual reality (Simeone, Speicher, Molnar, Wilde, and Daiber, 2019).

<sup>7</sup>See Wilde, Ballesteros-Mesa, and León Urrutia (2016a); Wilde, León Urrutia, and White (2016b); Wilde (2016); and Wilde, Urrutia, and Borthwick (2017).

<sup>8</sup>As reflected in my upgrade report, edited as a book chapter (Wilde and Zaluska, 2016).

<sup>9</sup>In my 2011/2012 classes I prepared some multiple-choice questions to students for informal assessment within lectures, to gauge whether they were facing some stumbling blocks before moving on to



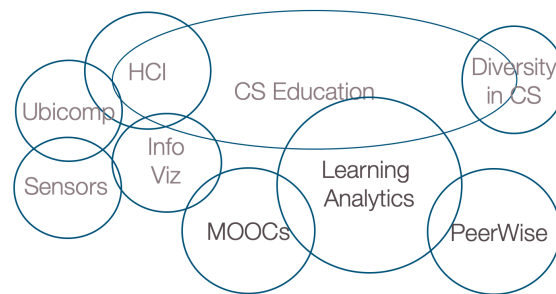


FIGURE 1.1: Computer Science Education (CSE) has been the main theme explored during my candidature. This thesis' focus is on learning analytics with Massive Open Online Courses (MOOCs) and PeerWise.

and the introduction of video coursework<sup>10</sup>. The latter led to collaborations<sup>11</sup> with researchers in computer science education (CSE) and human-computer interaction (HCI), including a series of workshops on video targetted to each community.

Given that my research agenda encompassed the wide range of interests described above, one of my challenges has been to delimit the scope of this thesis sufficiently to evidence in depth the ways I have been able to extend the forefront of my discipline. In what remains of this chapter, I explain how this thesis provides such evidence, starting from the motivation for this research (in Section 1.1), the specific research questions I addressed and the framework I used in doing so (in Section 1.2). I finalise with the thesis' organisation as a roadmap based on its title (in Section 1.3).

## 1.1 Motivation for this thesis

This research is part of a wider exploration on learner engagement using digital technologies in peer-supported environments, including those in the context of online learning, of which massive open online courses (MOOCs) are an important class, but also other online learning that are designed to complement face-to-face instruction, such as with the web-based peer-learning software PeerWise<sup>12</sup>.

more complex topics (Wilde, 2014). This worked well in my largest classes (with more 80 students), but less so in the smaller ones (with less than twenty).

<sup>10</sup>I introduced videos for assessment in my *Interaction Design* classes of 2015/16 and 2016/17 (Snow and Wilde, 2017; Wilde and Snow, 2018a). Another innovation introduced in these classes was the use of PeerWise, explained in Chapter 6.

<sup>11</sup>Namely, Vasilchenko, Wilde, Snow, Balaam, and Devlin (2018); Wilde, Dix, Evans, Vasilchenko, Maguire, and Snow (2019). Also, a number of events were collaboratively organised too, such as the HCIvideoW workshop (Wilde and Dix, 2020a) and workshops on using video in computer science education (Wilde, Vasilchenko, and Dix, 2018; Wilde and Terzic, 2018; Wilde and Dix, 2020b).

<sup>12</sup>PeerWise: Ask | Share | Learn <https://peerwise.cs.auckland.ac.nz/> (Last accessed on 2<sup>nd</sup> December 2020).

Relatively speaking, there is not much published research on learning analytics on face-to-face instruction data. One of the reasons is arguably that the data has much greater variety than in MOOCs (and it is not as well structured either), where however, there is a lot of research that can be used to inform approaches to understand learner engagement in this space too.

Even within the MOOC space, where there is a fertile ground for learning analytics research, I can identify a gap lying at the little-explored intersection of *heuristic-based classification approaches*, which are interpretable but often rigid and biased to preconceptions about learner behaviours rather than being based on what learners actually do; and *unsupervised learning approaches*, which are able to elicit from the data what is actually happening in practice, but often do it through models of limited interpretability. In my review of existing studies to date to the best of found knowledge, that apply interpretable clustering methods on both online and face-to-face instruction.

Having identified these gaps, I hence have refined my wider interest into human-centric computing and learner success in general to focus on the study of an operationalisation of learner engagement that can apply to both kinds of learning environments. I do so by identifying some approaches in the literature about MOOCs that are applicable to other peer-supported digital environments, such as PeerWise, but also many others, by articulating a general model of learner interaction, and validating it by producing interpretable clusters of learner engagement in various contexts.

Prior the formulation of the research questions addressed in this thesis, it is essential to understand the population of interest (learners in peer-supported digital environments) and the behaviours being studied (their engagement within, or more precisely, proxies of their engagement). The following are the operational definitions of these terms in the context of this thesis that have emerged from the discussion on relevant literature around learning, learners and engagement, particularly within educational technologies that support peer-learning (in Section 2.1):

**Definition 1.1** (*Learners in a peer-supported digital environment*). All users of a peer-supported digital environment who interact with learning content therein available, and who are able to interact with other users within.

**Definition 1.2** (*Learner engagement in peer-supported digital environments*). All of the behavioural, cognitive or emotional interactions by learners within peer-supported digital environments.

**Definition 1.3** (*Proxies of learner engagement in peer-supported digital environments*). Digital (and therefore measurable) traces of behavioural, cognitive or emotional interactions by learners within said environments.

## 1.2 Purpose of this research

Certain types of engagement behaviour may manifest differently in the contexts of different platforms, and therefore be captured with different variables, that can be used as proxies of engagement. The aim of this thesis is therefore to answer the following four research questions:

- RQ1** How can learner engagement be meaningfully compared across peer-supported digital environments?
- RQ2** What does a data-driven approach to learner interactions reveal about learning engagement within FutureLearn MOOCs?
- RQ3** What does a data-driven approach to learner interactions reveal about learning engagement within the PeerWise digital environment for face-to-face instruction?
- RQ4** Is learner engagement different in different kinds of peer-supported digital environments, be it a complement to face-to-face instruction, or a fully online course?

These questions are addressed in this thesis following the research framework outlined in the diagram presented in Figure 1.2.

## 1.3 Thesis outline

This thesis, titled “*A Platform-Agnostic Model and Analysis of Learner Engagement within Peer-Supported Digital Environments: FutureLearn MOOCs and PeerWise*” is organised as follows:

Chapter 2 offers a literature review around the main concepts explored in this research, starting with learning and engagement (Section 2.1), including considerations on behaviour change, experimentation in educational research, learning at scale and peer-learning; some peer-supported environments, such as massive open online courses (MOOCs, in Section 2.2), in general but also FutureLearn in particular; followed by PeerWise (Section 2.3) and other peer-supported digital learning environments (Section 2.4). Then I look into techniques for the measuring, collection and analysis of these (part of “learning analytics”, in Section 2.5), including considerations about feature engineering and unsupervised learning algorithms, clustering in particular. I also offer some definitions on measures for information retrieval (Section 2.7) and around Allen’s interval algebra (Section 2.6), both which are helpful to refer to later on.

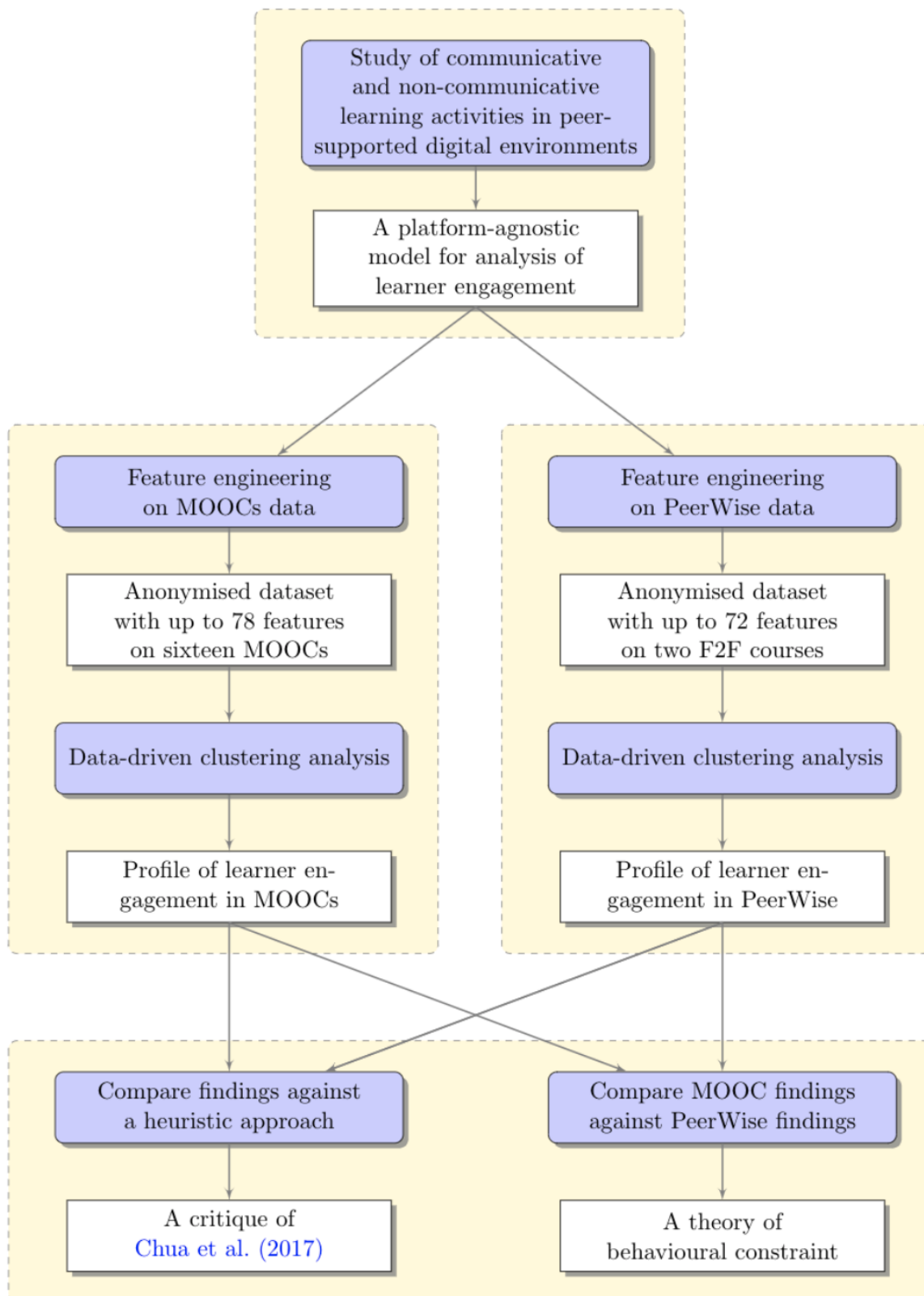


FIGURE 1.2: Research Framework Diagram, showing the processes (in purple, rounded boxes), the outputs (in white, squared boxes) for each of the four research questions of this thesis (in yellow, dashed boxes).

Chapter 3 outlines the methodology used throughout this thesis, both at high level and in detail, through a data science pipeline. Chapter 4, in offering an answer to **RQ1**, presents a formalism for a platform-agnostic model of learner interactions within peer-supported digital environments. I use this model to inform data-driven analyses of interactions in two very different environments, each described in the following two chapters: Chapter 5, “Peer-learning online within FutureLearn MOOCs” and Chapter 6, “Peer-learning in face-to-face instruction mediated by PeerWise”. In those chapters I answer research questions **RQ2** and **RQ3**, as I present the results of my analysis of digital traces of activity captured within each of those peer-supported digital environments. I do this through a data-driven approach, specifically with unsupervised learning, using a feature engineering process informed by the model earlier formalised. In the case of MOOCs, I compare this data-driven approach against a heuristic-based approach reported in the literature. I then compare these environments to each other given the findings of the previous two research questions to articulate answers to **RQ4**.

Finally, the conclusions are presented in Chapter 7, as well as the limitations of my research and pointers to future work.

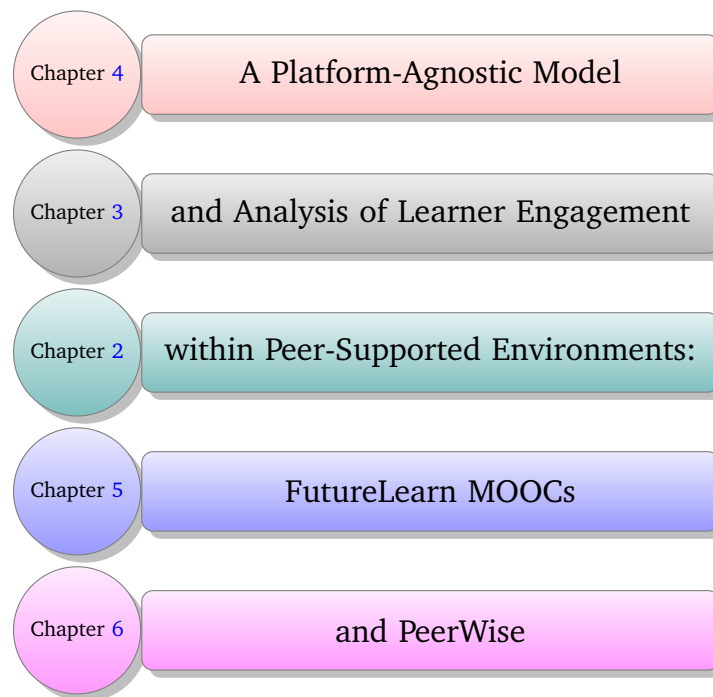


FIGURE 1.3: Organisation of this document, linking the contents of each chapter to the topics in the title of this thesis.



## Literature Review

*May all your problems be technical ones*

Old programmers' blessing (according to Ian Witten and David Bainbridge in "A RETROSPECTIVE LOOK AT GREENSTONE: LESSONS FROM THE FIRST DECADE", *ACM/IEEE Digital Libraries*, pp. 147-156. June 2007.)

Many of the challenges encountered in this thesis echo in nature those reported by [Witten and Bainbridge \(2007\)](#) when they reflected on the first ten years of the Greenstone project. They mentioned "political, educational, and sociological" challenges, which prompted them to recall the old programmers' blessing quoted in the epigraph. The literature reviewed in this area covers some of these types of challenges, made evident through the impact of digital technologies in the provision of support and feedback to learners and other stakeholders of educational institutions. In my review I consider aspects of learning in peer-supported environments (MOOCs and PeerWise specifically but also others in general) as well as those needed for characterising the students via learning analytics. This makes possible the identification of behaviours to better understand learners and their engagement.

This chapter is organised as follows: Section [2.1](#) defines engagement in the context of learning and it looks at the related problems of behavioural interventions to increase it, as well as experimental constraints within educational environments and within peer-learning technologies specifically. These lead to Sections [2.2](#) and [2.3](#), where two examples of peer-supported environments are given: FutureLearn MOOCs and PeerWise. Some others are covered in Section [2.4](#). Then I give a whistle-stop tour around the fields of educational data mining and learning analytics in Section [2.5](#), paying special

attention to feature engineering and clustering, as they are particularly relevant to this thesis. In Section 2.6, I give some definitions on interval algebra (upon which rest some of the formalisms I present in Chapter 4). Finally, Section 2.7 gives some information retrieval definitions that are useful in appraising the results of the experiments I present later, in Chapters 5 and 6.

## 2.1 Learners and engagement

Though the terms *learner* and *engagement* are widely used in a variety of contexts, it is worth to discuss the lack of an authoritative consensus on how they are understood. The layperson's accepted view of learning encompasses a change of behaviours or attitudes upon sufficient exposure to knowledge or practical skills (often, but not exclusively, through teaching). This interpretation is echoed in recent literature (Darling-Hammond, Flook, Cook-Harvey, Barron, and Osher, 2020) though these authors' research focuses in schoolchildren. Such a focus on the school context is also seen in most of the research since the early twentieth century, as observed by Laurillard (2013) and Biggs and Tang (2007). More recently however, the interest on higher and post-compulsory education has grown, together with innovations in pedagogy, facilitating updates on conversations around learning. These include wider considerations such as the use of technology to mediate the process of acquisition of knowledge and skills, but also less tangible aspects such as experiences of flow and personal wellbeing (Kukulska-Hulme, Bossu, Coughlan, Ferguson, FitzGerald, Gaved, Herodotou, Rienties, Sargent, Scanlon, Tang, Wang, Whitelock, and Zhang, 2021).

Consequently, the term *learner* has also suffered from this lack of clarity, yet it is also commonly used and rarely explicitly defined. Whilst often used interchangeably with the term *student*, the term "learner" should be understood as a more generic, characterising someone engaged in a learning process, whereas the term "student" is more specific: used for learners who are taught. This differentiation allows for a discussion of learner behaviour (and particularly engagement, as discussed below) independently of whether there is a teacher, or even whether the learning is taking place in a non-formal, informal or formal context (see Tudor (2013) for an in-depth discussion of the differences between these).

Furthermore, accepting a learner to be anyone involved in learning allows for teachers, instructors and facilitators to also be considered as learners since they are co-participants of the process. This is understood well by Paulo Freire (1970) when he celebrated students and teachers coming together as equal learners. As part of his



pedagogy, the teacher is not considered to have the monopoly on expert knowledge. Learning is seen as a liberating experience for all parties (in contrast with a “banking” experience, where knowledge is imparted by teachers and “saved” by students), and this liberation is through dialogue. Kolb (1998) concurs, not without indicating that “dialogue among equals doesn’t mean that in any single conversation there isn’t a point in which one person is an expert and the other person is not.”

The term *engagement* has been shown to be equally challenging to define precisely, despite its wide adoption and relevance within learning contexts. Indeed, learner engagement is considered one of the primary models applied to understand dropout and fostering completion, as noted by Reschly and Christenson (2012). However, in the context of peer-supported digital environments, such as MOOCs, no definition has been widely adopted by the community. Gore (2018) had researched extensively on this topic, and I quote (the italics are mine):

“In reviewing academic papers relating to engagement, *very few* had an actual definition of the term within them (Cormier and Siemens, 2010), and *none* addressed the context of learning of MOOCs, with most relating to the traditional classroom setting (Becker, 2000; Kuh, 2001; Kuh and Gonyea, 2003; Ahn et al., 2013; Milligan et al., 2013; Ramesh et al., 2013).”

More recently, Maia, Araújo, Figueiredo, and Serey (2020) recognised that learner engagement is an essential aspect of learning, involving behavioural, cognitive and emotional processes. As such, it can be operationalised in many ways, depending on factors such as the pedagogy behind the learning design. One important aspect of learning design, discussed in Chapter 1, is the transformation that the traditional lecture format is seeing into the adoption of an increasingly learner-centric model. Though much criticised nowadays, Prensky (2001) highlighted how tertiary instructors worldwide have been faced with the problem of how best to teach and engage the current generation of “digital-native” students who arguably display a decreased tolerance to traditional teacher-centric lecture style information dissemination. Whilst the validity of such a categorisation of students based on generational traits has been widely discredited (Kennedy, Judd, Dalgarno, and Waycott, 2010; Hockly, 2011), some of the implications from Prensky’s views have been very influential (Palfrey and Gasser, 2010; Jones, 2011a; White, Connaway, Lanclos, Le Cornu, and Hood, 2012).

Educational technology and the teachers using it do face the challenge of keeping apace of emerging technology and leveraging it to increase engagement and maximise learning outcomes in the classroom (Tondeur, van Braak, Siddiq, and Scherer, 2016).

As a response to these learning aspirations, a growing number of university instructors seek to provide students with more direct input in their learning process.

The peer-learning model re-positions the instructor as a facilitator rather than a sage, even as a learner, as Freire (1970) had envisaged. Peer-learning has been credited with realising a greater level of productivity and learner engagement than traditional content delivery (Unruh, Peters, and Willis, 2016). This model can be facilitated through the use of online resources and teaching software that allow conversations and coursework to extend outside of class time (Mehring, 2016), often bundled and accessed through one single portal, as in the examples discussed in Sections 2.2, 2.3 and 2.4. These are collectively known as “peer-supported digital environments”.

The above discussion is the basis of the operational definitions given in the introduction, in particular, Definition 1.1 (learners in peer-supported digital environments), Definition 1.2 (their engagement within) and Definition 1.3 (proxies of engagement).

I next consider how to increase engagement (subsection 2.1.1), and conduct experimentation in this context (subsection 2.1.2) and the particular challenges of learning at scale (subsection 2.1.3). Then I return to peer-learning technologies (in subsection 2.1.4) to complete this section about learners and engagement.

### 2.1.1 Behaviour change for engagement

Learners make behavioural choices (with various degrees of intentionality) as they engage within learning environments. They do so as they are exposed to information about their past engagement and that of their peers as a whole, or even just a handful of “successful” peers, such as for example, those at the top of a leaderboard in a gamified environment. This information, amongst several kinds of ‘nudges’, are often used by platform designers as these are understood to be helpful in increasing learner engagement.

In the context of behavioural interventions, the term *nudge*, as used by Balebako, León, Almuhimedi, Kelley, Mugan, Acquisti, Cranor, and Sadeh (2011) and Acquisti (2009) was first introduced by Thaler and Sunstein (2008) to describe “any aspect of the choice architecture that alters people’s behaviour in a predictable way without forbidding any options or significantly changing their economic incentives.” By *choice architecture* these authors refer to the environment (either social or physical) in which individuals make choices. There is an element of low-awareness on the part of the individual of such an architecture, so the individuals are still exercising their free will when making choices, however such a choice might have been different were it not for the

intervention. A taxonomy of different types of “behaviour change interventions” ([Great Britain. Parliament. House of Lords, 2011](#)), including examples, is presented in [Table 2.1](#). In this table there are possible nudges that platforms for peer-supported digital environments may include in order to increase engagement, more particularly so, those within the last intervention category: “Guide and enable choice”, in particular:

- *Persuasion*: By directly encouraging learners to engage in behaviours.
- *Provision of information*: By raising awareness of own behaviours through a summary of past engagement, reaching specific milestones, as well as peers’ interactions on past contributions.
- *Use of social norms and salience*: By providing information about the peers’ engagement, such as through leaderboards or other visualisations.

It is possible, therefore, to “nudge” (in Thaler and Sunstein’s sense) learners into behaviours of higher levels of engagement, and various platforms do these in different ways.

## 2.1.2 Experimenting in educational research

As [Cohen, Manion, and Morrison \(2007\)](#) points out, in educational research it is often the case that true experimental design in the strict sense of the word is not possible, given that they cannot be conducted under laboratory conditions where all variables can be controlled, or it is not possible to apply controls typically used in field experiments. Further, in many of these cases, exact repeatability is challenging or even impossible. [Cohen et al. \(2007\)](#) covers a number of research methods that are appropriate to educational research, amongst which, two are relevant to this thesis: quasi-experiments and ex post facto research.

A *quasi-experiment* is an empirical study of the causal impact of an intervention on a target population without random assignment. Quasi-experiments are commonly used in education and other disciplines where it is not practical, ethical or reasonable to randomize study participants to the treatment condition ([Cohen et al., 2007](#)). One of the most widely applied types of such quasi-experiments is the non-equivalent control group design, by which the two groups (the “experimental” and “control” groups) are non-equivalent in the sense of not having been drawn by randomisation, and therefore may be subject to uncontrolled variables unevenly.

TABLE 2.1: Table of interventions. (Adapted from [Great Britain. Parliament. House of Lords \(2011\).](#))

Regulation of the individual	Financial measures directed at the individual	Non-regulatory, non-financial measures in relation to the individual									
		Choice Architecture (“Nudges”)									
Guide and enable choice											
Category	Eliminate choice	Restrict choice	Financial disincentives	Financial incentives	Non-financial incentives and disincentives	Persuasion	Provision of information	Changes to physical environment	Changes to the default policy	Use of social norms and salience	Examples
	Prohibiting goods or services	Restricting the options available to individuals	Fiscal policies to make behaviours more costly	Fiscal policies to make behaviours financially beneficial	Policies to reward or penalise behaviours	Persuading individuals using arguments	Providing information in leaflets	Altering the environment	Changing the default option	Providing information about what others are doing	

*Ex post facto* research is done retrospectively ('after the fact', as translated from Latin), and but the dependent variables are examined in retrospect for their possible relationship to test hypotheses about cause and effect on the independent variable. It can be used to study groups that are already different in some respect and search in retrospect for the factor that brought about the difference (Cohen et al., 2007).

### 2.1.3 e-Learning and Learning at Scale

Access to higher education has increasingly widened in recent years, with greater expectation for school leavers to pursue further studies in the hope of increasing their chances for employability and social mobility. For some degrees, such as Computer Science, there has been a growth in student numbers, resulting in many classes that comprise learners in their hundreds. This trend is observed in higher education institutions across the UK<sup>1</sup>. Such an unprecedented growth has meant that higher education institutions must adapt in order to rise to challenges in assessment and feedback, with views to improve sustainability and scalability, all the while serving the primary goal of facilitating learning.

Parallel to this thought in the face-to-face instruction space, we have the issue of an ever growing affordability of smartphones, portable computers, and the ubiquity of the Internet, which not only allows students to access learning materials "anytime and anywhere", but also facilitates the uptake of online learning. Indeed, a natural consequence of the pervasiveness of digital technologies in recent years is that they are now almost universally used in teaching and learning (to various degrees). Indeed, they have been intertwined for a long time. Coinciding with the advent of the personal computer in the 1970s, the term *Computer Assisted Learning* was first coined, alongside *Computer Assisted Instruction* and similar others, however, these terms are less commonly used as they are becoming replaced in the educational discourse by the term *e-learning*. The former have been used to characterise the use of computers in education, or more specifically, where digital content is used in teaching and learning. In contrast, the latter is generally used only when the content is accessed over the Internet (Derntl, 2005; Hughes, 2007; Jones, 2011b; Sun, Tsai, Finger, Chen, and Yeh, 2008). Salmon's model of online learning (Salmon, 2002) is represented by a five-step hierarchy with increasing levels of

---

<sup>1</sup>Computer science experienced the "largest percentage increase" (of 4%) in enrolments of first year undergraduate students between 2015 and 2017, according to the Higher Education Statistics Agency (HESA, <https://www.hesa.ac.uk/news/11-01-2018/sfr247-higher-education-student-statistics/subjects>). More recently, Computer Science continues to increase enrolments, to 8% of all first year students choosing this subject between 2017 and 2019 (<https://www.hesa.ac.uk/news/16-01-2020/sb255-higher-education-student-statistics/subjects>). (Pages accessed 31<sup>st</sup> March 2020).

interactivity though counting with visible e-moderating and technical support. In this context she used the term e-tivities to mean a framework for designing learning activities by individuals and groups, typically engaged asynchronously with the learning.

Not restricted to online environments, *learning at scale* is the study of the technologies, pedagogies, analyses, and theories of learning and teaching that take place with a large number of learners and a high ratio of learners to facilitators. The scale of these environments changes the very nature of the interaction and learning experiences. The impact of learning at scale can be seen in different areas, but one in which it is most evident is in the increased complexity of data at scale. Many online learning environments keep traces of learner interactions, as well as their engagement and performance. This is typically kept in heterogeneous and distributed systems which make their processing and interpretation challenging. In fact this is much more complex for institutional data in the face-to-face context than in online learning, because the scale in the latter context is given by the number of learners, but the actual data tends to be somewhat standardised and centralised in a way that is not necessarily possible in the former (Dix, 2016).

### The 90-9-1 principle of online engagement

Noteworthy amongst the phenomena that can be observable in learning at scale is the application of the 90-9-1 principle, which is attributed to be the common distribution of engagement in Internet communities (Carron-Arthur, Cunningham, and Griffiths, 2014). It was first observed by van Mierlo, Voci, Lee, Fournier, and Selby (2012) in the context of social networks for smoking cessation. Essentially, this principle is a variation of the Pareto principle (where 20% of a group will produce 80% of the activity). The 90-9-1 rule posits that in a collaborative environment online, 90% of the participants are 'lurkers', who watch but not contribute, 9% make changes or updates, and 1% add new content. Hence, theoretically one might expect that in these learning environments 90% are asocial, 1% are fully engaged, and the remaining 9% exhibit a more passive engagement somewhere in between both extremes.

#### 2.1.4 Peer-learning: Experiential and conversational learning

The dialectic between individual action and co-reflection has tensions as per these two complementary views: For one, the learning journey is *individual-centered* (and hence the acquiring of skills is through "doing" and individual's reflections on how this is done). For the other, conversations with *others* enable this co-reflection and sharing of knowledge which otherwise would not change the individual.

Pedagogic conceptions of teaching and learning are usually understood in the literature as falling into one of two categories: teacher-centred (content driven) and student-centred (learning driven) (Jones, 2011b, and references therein). Figure 2.1 shows these orientations as overarching the main five conceptions of teaching and learning which act as landmarks alongside a continuum of roles in learning. Deep learning occurs at the bottom end of the scale, as opposed to shallow learning which occurs at the top end. When student-centred, computer assisted learning can increase students' satisfaction and therefore engagement and attainment. It is remarkable that the move towards learner-centredness in Higher Education coincides with the trends towards personalisation and user-centredness in Human-Computer Interaction and computing technologies in general.

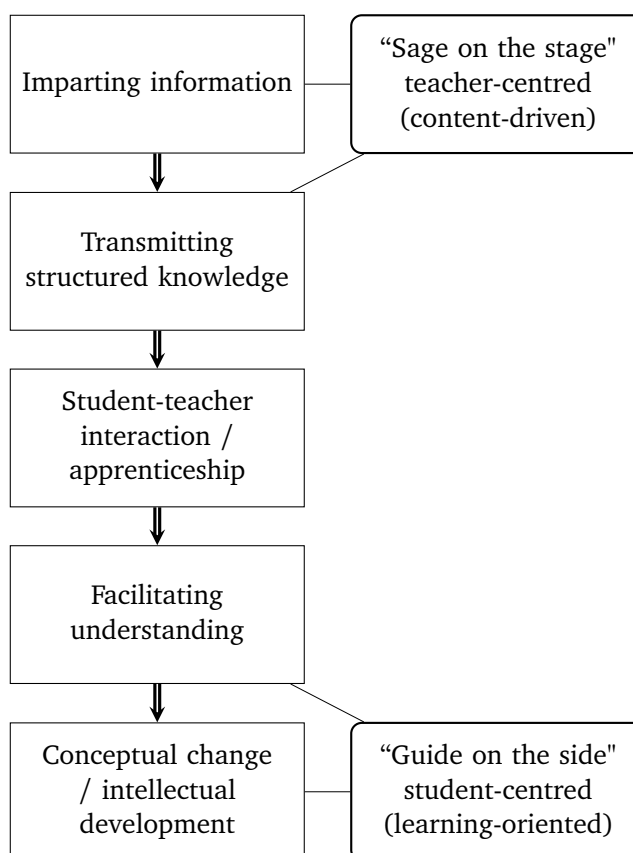


FIGURE 2.1: Multi-level categorisation model of conceptions of teaching: the teacher progresses from “sage on the stage” to “guide on the side” (adapted from Kember (1997).)

The trend towards a widespread use of mobile devices, earlier identified, brings an increased number of opportunities to effect the *conceptual change* from the categorisation above, as it has the potential of making the learning more student-centred than before: it would take place wherever the student goes, whenever it suits the student

best<sup>2</sup>. The advent of the web has made online learning accessible for all (Bates, 2005), which has undoubtedly transformed access to information and knowledge in the last twenty years, through distance learning and Massive Open Online Courses (MOOCs) (Yuan et al., 2013).

Additional opportunities to reach students to either deliver content or to assess their learning, are coupled with opportunities for other stakeholders at educational institutions to gain an insight on student achievement (typically progression and completion) via learning analytics, as presented in the next section.

Peer-learning technologies offer rich opportunities for extending learning beyond the classroom. Students may work cooperatively in the formulation and peer-assessment of multiple-choice questions.

## 2.2 Massive Open Online Courses

Nowadays there is an abundance of opportunities to study learning phenomena of the kinds described in section 2.1. These opportunities have arisen due to the sustained proliferation of Massive Open Online Courses (MOOCs) around the globe ever since its emergence in the late 2000s, but more dramatically since 2012, the often called “year of the MOOC” (Pappano, 2012). This trend is shown in Figure 2.2, where there are over 16,000 courses as of now, according to (Shah, 2020b).

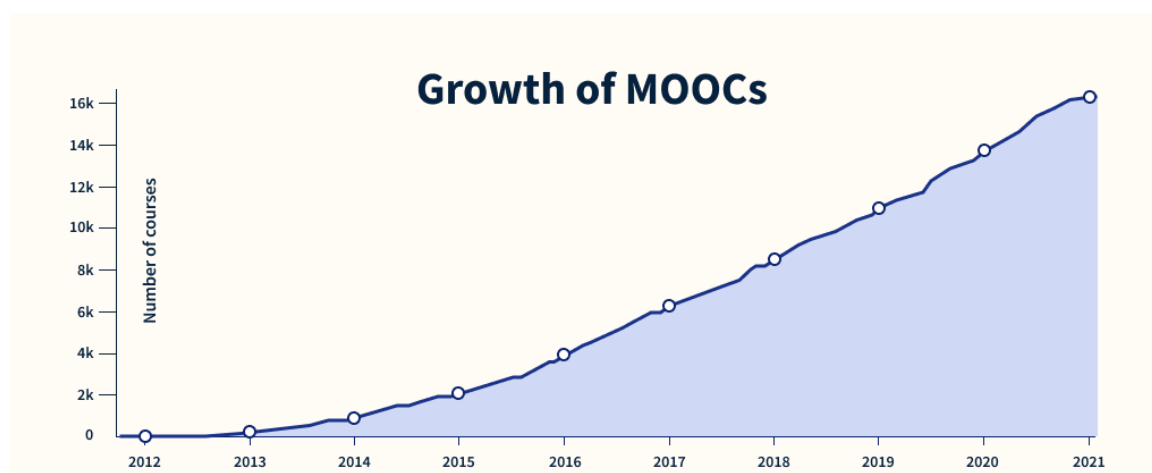


FIGURE 2.2: Growth of MOOCs since 2012 as reported on Class Central. Statistics do not include China. (Shah, 2020b)

<sup>2</sup>The “anywhere, anytime” maxim driving pervasive computing is also a motivator for the development of the next generation of e-learning. Rubens, Kaplan, and Okamoto (2014) discuss the evolution of the field, aligning it to the advent of Web 2.0 and 3.0, central to this paradigm of learning.



The emergence of MOOCs is a consequence of the increased interconnectivity of the digital age. When [Siemens \(2005\)](#) proposed connectivism as a new theory to sit alongside classical learning theories (of which Piaget's constructivism is an example [Fox \(2001\)](#)), pioneer online courses started to be created based on this theory: people learn by making meaningful connections between knowledge, information resources and ideas during the learning process. The key to a successful connectivist course would therefore be the use of a platform which fosters the formation of such connections in a distributed manner. These courses have become known as c-MOOCs, of which the first one was delivered in 2008 by Siemens and Downes ([McAuley, Stewart, Siemens, Cormier, and Commons, 2010](#); [Rodriguez, 2012](#)).

TABLE 2.2: Summary of similarities and differences between c-MOOCs and x-MOOCs.  
Adapted from [Cobos et al. \(2017\)](#)

Characteristic	c-MOOCs	x-MOOCs
Number of learners	Should scale to large numbers	Should scale to large numbers
Method of delivery	Online	Online
First MOOC delivered (year)	Connectivism and Connective Knowledge (2008)	Introduction to Open Education (2007)
Related learning theory	Connectivism	Instructivism or behaviourism
Communication	Distributed	Centralised
Design mainly supports	connections between learners, resources and ideas (connections with knowledge objects)	relationships between teachers and learners, mediated through task completion (personalised interactions with instructional materials)

In contrast, other courses were designed to adapt the medium, learning materials and assessments of traditional courses (i.e. instructivist, or cognitive behaviourist courses (Anders, 2015)) so that these could be delivered at scale. Under instructivism, learning is also an active process, but the relationship between teachers and learners is key – the relationship is mediated through specific tasks which are assessed as a measure of the learning process. These MOOCs have become known as x-MOOCs, a term coined by Downes in 2012 to differentiate them from his c-MOOCs. The first x-MOOC was delivered in 2007 though: the Introduction to Open Education, by David Wiley from Utah (Fini, Formiconi, Giorni, Pirruccello, Spadavecchia, and Zibordi, 2008). Noting that there are many similarities as well as important differences between c-MOOCs and x-MOOCs<sup>3</sup> (summarised in Table 2.2), it is interesting to consider them when analysing case study courses in practice. The inherent similarities and differences between these models (which in turns translate in a number of affordances provided by MOOC platforms) may influence learner behaviours.

A scoping review of the literature in MOOCs produced Table 2.3, paying particular attention to methodologies and outcomes by research seeking to categorise learners.

---

<sup>3</sup>Cobos et al. (2017) could be forgiven for having considered FutureLearn MOOCs as c-MOOCs, given that the platform is not founded on instructivism but on a learning pedagogy as conversation, as discussed in subsection 2.2.1. But they are not c-MOOCs either, as they have a clear, fixed structure in each offering, and indeed it can be argued that there are no longer any true c-MOOCs.

TABLE 2.3: Summary of reviewed categorisations of MOOC learners in the literature

Authors	Data description	Methodology	Categories (number of learners)
Milligan, Littlejohn, and Margrayn (2013)	Type: c-MOOC Course: #change11 Weeks: 35 Learners: 2,300 (+ 3 facilitators) Interviewed participants: 29 (of 35 who had agreed to take part).	Qualitative (interviews). Explored how learners followed the course, whether they contributed in blogs (and on social media) and whether they persevered.	<ul style="list-style-type: none"> <li>· Active participant (12)</li> <li>· Lurker (13)</li> <li>· Passive participant (4)</li> </ul>
Kizilcec, Piech, and Schneider (2013)	Type of MOOCs: x-MOOCs Platforms: edX and Coursera Courses: <i>Computer Science 101</i> (HS level), <i>Algorithms: Design and Analysis</i> (UG level) and <i>Probabilistic Graphical Models</i> (GS level). Assessment periods: 9 for each course Active learners: 46,096 (HS), 26,887 (UG) and 21,108 (GS).	Quantitative. A clustering analysis of learners' sub-populations (using k-Means) based on individual engagement per assessment period, as being: <ul style="list-style-type: none"> <li>· On track,</li> <li>· behind,</li> <li>· auditing or</li> <li>· out.</li> </ul>	<ul style="list-style-type: none"> <li>· Auditing (1,486)</li> <li>· Completing (2,352)</li> <li>· Disengaging (739)</li> <li>· Sampling (301)</li> </ul>
Alario-Hoyos, Pérez-Sanagustín, Delgado-Kloos, G., and Muñoz-Organero (2014)	Type: x-MOOC, assessed Platform: MiriadaX Course: <i>Digital Education of the Future</i> weeks: 9 (+ Easter break) learners: 1,951 (+ 5 facilitators) comments: 10,033 (+ 363 from facilitators) steps: not studied	Quantitative, based on engagement in tasks set by facilitators (including peer-reviewing) as well as their social media usage.	Three categories with patterns: <ul style="list-style-type: none"> <li>· Lurkers: No shows (1,486) and Ob-servers (2,352)</li> <li>· Participants who do not complete the course: Drop-ins (739), Latecomers (301) and Drop-in latecomers (292)</li> <li>· Participants who complete the course: Non-engaged (88) and Engaged (337).</li> </ul>
Anderson, Huttenlocher, Kleinberg, and Leskovec (2014)	Type: x-MOOC Platform: Coursera Courses: <i>Machine Learning</i> (3 runs), <i>Probabilistic Graphical Models</i> (3 runs) weeks: learners: comments: steps:	Quantitative. They defined formulas with participation thresholds for each category and then studied what other factors may correlate with their behaviour. They found that the time of interaction (relative to registration and assignments dates) has an effect on the learner category.	<ul style="list-style-type: none"> <li>· Bystanders (41-62%)</li> <li>· Viewers (18-25%)</li> <li>· Collectors (11-21%)</li> <li>· All-rounders (4-20%)</li> <li>· Solvers (0-1%)</li> <li>· Archaeologists (18%, orthogonal to the ones above).</li> </ul>

Continued on next page

Table 2.3 – Continued from previous page

Authors	Data description -	Methodology	Categories (number of learners)
Ferguson and Clow (2015b,a)	Type: x-MOOC Platform: FutureLearn Courses: four Open University early courses in Physical Sciences, Life Sciences, Arts, and Business respectively weeks: 6-8 weeks learners: from the UK in its majority (numbers not reported) comments/steps: not reported.	Quantitative. k-Means clustering (with silhouette method to determine optimal number of clusters), intending to replicate Kizilcec et al. (2013). They found new clusters, which were not present across all their observed courses. Learner behaviour varied across courses, with one MOOC being dominated by Samplers and another one having 2-3 times the number of keen completers than the others.	· Samplers (37-56%) · Strong starters (8-14%) · Returners (6-8%) · Midway dropouts (0-8%) · Nearly there (5-6%) · Late completers (0.2-8%) · Keen completers (7-23%) · three other supplementary clusters (0 - 20%).
Tseng, Tsao, Yu, Chan, and Lai (2016)	Type: x-MOOC Courses (and students): five YZU courses in C# programming (400), Internationalisation strategy (340), Computer-aided Design and Manufacture (749), Electronics (193) and English for Engineering (381). 1,489 students in total Year: 2014 Weeks: 9-19 weeks Learners: 94% Taiwanese	Quantitative. k-Means clustering (with hierarchical clustering's Ward method to determine optimal value of k). Learner participation in discussion fora was very similar across courses, with the exception of C# programming, which had more active and passive learners and fewer bystanders. The bystanders had much lower completion rate than any other group.	· Active learners (1%) · Passive learners (9%) · Bystanders (90%).
Sunar, White, Abdullab, and Davis (2017)	Type: xMOOC Platform: FutureLearn Course: <i>Developing Your Research Project</i> Weeks: 8 Learners: 9,855 enrolled, 2,927 active Comments: unspecified (reportedly the highest number amongst Southampton MOOCs in 2014).	Quantitative. Through descriptive statistics, investigating the relationship between posting behaviours against following behaviours.	· posters who did not follow anyone (1316) · followers who post (551) · followers who do not post (238).

Continued on next page

Table 2.3 – Continued from previous page

Authors	Data description -	Methodology	Categories (number of learners)
<a href="#">Chua et al. (2017)</a>	Type: x-MOOC Platform: FutureLearn Course: <i>Inequalities in Personal Finance</i> weeks: 4 (+ 2 of logged data beyond end of the course) learners: 1,951 (+ 4 facilitators) comments: 10,033 (+ 363 from facilitators) steps: not studied	Quantitative, based on whether they had contributed any comments of the types: · Initiating post, · lone post, · reply, · further reply and · initiator's reply.	Social learners (936) of types: · Loner (164) · Replier (60) · Initiator without replying (114) · Initiator who responds (37) · Active social learner (91) · Active social learner without turning taking (75) · Reluctant active social learner (5)
<a href="#">Dowell, Poquet, and Brooks (2018)</a>	Type: x-MOOC Platform: Coursera Course: unspecified Year: 2013 weeks: 8 learners: 644 contributions: 2,335 posts + 1,437 comments conversations (threads): 180 steps: not studied	Quantitative. A k-Means cluster analysis to discover communication patterns.	
<a href="#">Tubman, Oztok, and Benachour (2019)</a>	Type: x-MOOC Platform: FutureLearn Course: unspecified learners: unspecified conversations: 257,239 surveyed participants: 304 responses.	Quantitative. Learner engagement is measured with and without a visualisation plugin to extend interactive affordances in FutureLearn. Uses the taxonomy of comments by <a href="#">Chua et al. (2017)</a> and a survey on the perceived affordances of the platform. Learners valuing social interactions were more likely to: · respond positively to thinking through discovery and connecting new ideas; · find the visualisation useful; · comment more.	Learners are not categorised but conversational units, of types: · Extended social conversations, · Q & A, · Limited social conversations, and · Lone.

Continued on next page

Table 2.3 – Continued from previous page

Authors	Data description -	Methodology	Categories (number of learners)
Reich and Ruipérez-Valiente (2019)	Type: xMOOC Platform: edX Courses: 565 runs of 261 courses (from MITx and HarvardX) learners: 12.67 million enrolled, 5.63 million active activities: over 4.4 billion events.	Quantitative, processing clickstream events with the <i>edx2bigquery</i> framework to build up a person-course dataset with over 60 variables that describe the interaction of the student with every course they enrolled to.	<ul style="list-style-type: none"> <li>· Unique learners (0)</li> <li>· Participants (avg. 48.02%)</li> <li>· Explorers (avg. 25.59%)</li> <li>· Completers (unspecified)</li> <li>· Certified (avg. 5.03%).</li> </ul>
Kizilcec and Chen (2020)	Platform: Shupavu 219 (an SMS-based mobile learning course for children in grades 6, 9 and 12 in Kenya) Subjects: “The Covenant”, “Body Systems”, “Fractions” and “Numbers” learners: 93,819 duration: Unspecified – but only the first ten days of activity are analysed activities: 28,410,376 platform actions, including 1,515,550 quiz activities.	Quantitative (adapting their approach from Kizilcec et al. (2013)), a clustering analysis of learners’ sub-populations (using k-Means) based on individual engagement over the first ten days of use of the platform, having noted that attrition is extremely high from that day onward. <i>k</i> was chosen as three based on the Elbow method. Clusters semantics were interpreted by inspection of the cluster centroids.	<ul style="list-style-type: none"> <li>· Low activity (64,403 students)</li> <li>· Medium Activity (11,120 students)</li> <li>· High Activity (1,814 students).</li> </ul>
Sunar, Abbasi, Davis, White, and Aljohani (2020)	Same as Sunar et al. (2017). This is a follow-up study using the same data.	Quantitative, with descriptive statistics based on heuristics emerging from data inspection. Though “clusters” are mentioned, these were not found through unsupervised learning. Several orthogonal categories were explored: - Based on the step completion · Inactive, Very rare, Rare, Moderate, Frequent and High (unspecified numbers of learners for each) - Based on the length of “chains” of comments - Based on the timing of comments.	<ul style="list-style-type: none"> <li>· Simple (717, 36.6% of social learners)</li> <li>· Moderately frequent (171, 8.7%)</li> <li>· Frequent (976, 49.5%)</li> <li>· Persistent frequent (107, 5.4%)</li> <li>· One-week contributors (717, 36.6% of social learners)</li> <li>· Continuous passive contributors (717, 36.6% of social learners)</li> <li>· Continuous active contributors (717, 36.6% of social learners).</li> </ul>

Continued on next page

Table 2.3 – Continued from previous page

Authors	Data description -	Methodology	Categories (number of learners)
Poquet, Jo- vanovic, and Dawson (2020)	Type: x-MOOC Platform: edX Courses: Two in engineering, one in data analysis and one in computer science Years: 2013-2014 Weeks: 8-11 Learners: $M = 40878$ , $SD = 10972$ contributions: 16,195 posts over 4,260 discussion threads conversations (threads): 180 steps: not studied	Quantitative, using agglomerative hierarchical clustering patterns of change in communication, which were identified by coding on whether the comment was on Content Task (CT), content non-task (CN), social or informational queries. One important finding is that contributions made by residents have richer variety of types of posts. Forum presence was found to be associated with the rate of posting in social non-task discussions, which decreased towards the end of the course.	Seven clusters, which are interpreted as variations of only two distinguished classes of learners: Residents and visitors.
Dowell and Poquet (2021)	Type: x-MOOC Platform: Coursera Course: unspecified, presumably the same as in Dowell et al. (2018) Year: 2021 weeks: 8 learners: 644 contributions: 2,335 posts + 1,437 comments conversations (threads): 180 steps: not studied	Quantitative, combining Group Communication Analysis (GCA), based on temporal semantic properties of online discourse, and Social Network Analysis (SNA) reflecting structural interpersonal patterns of online interactions.	Lurkers (170), Followers (69), Socially Detached (55), Influential Actors (168) and Hyper Posters.



From the research outlined in Table 2.3 I provide some additional details:

For example, [Littlejohn, Milligan, and Margarayn \(2011\)](#) identified workplace behaviours that are applicable to learning. These are identified as: *consume*, *connect*, *create* and *contribute*. In a survey of 462 respondents, they found that the combination of these behaviours helped them in their personal learning and work environments, as manifested in many learning practices. For example in formal learning (consuming and connecting), in teaching others (creating and contributing), and in learning through experience (all four). They conclude that technology tools for connecting may be most effective when they interface with tools for consuming. Similarly technology tools for creating knowledge resources may be most effective when they interface with tools for contributing knowledge to the collective”.

[Anderson et al. \(2014\)](#) were the first (amongst my reviewed literature) to report insights on the effect of the time of interaction (relative to registration and assignments dates) on the learner categories determined by the volume of interactions. For example, students who enrolled as early as six months in advance in the Coursera MOOCs they studied tend to become “bystanders” (70%) whereas those who enrol around the formal start of the course are much less likely to do so (35%). Also, a sizeable fraction of learners (18%) engaged in their first interaction after the course had finished (hence dubbed “archaeologists” in their taxonomy), and that only approximately 60% had enrolled prior to the formal start of the course. Another interesting observation from their data exploration was the fact that learners achieving the top marks in their assessments consumed the most lectures (with the majority watching videos more than once). However, those under the “solvers” category did not. They were dubbed “solvers” if the fraction  $a/(a + l)$  was above a certain threshold, given the number of assessments they engaged in ( $a$ ), and the number of lecture videos watched ( $l$ ). With regards to commenting behaviour, the authors looked into the difference in marks achieved by thread initiators and those who later contribute to the thread, with no generalisable findings.

[Reich and Ruipérez-Valiente \(2019\)](#) studied all the MOOCs delivered on edX by MITx and HarvardX from the start of the initiative in 2012 to 2018, comprising 565 course iterations from 261 different courses that have summed more than 12.67 million course registrations from over 5.63 million learners that generated over 4.4 billion events and invested more than 48 million hours in such courses.

[Sunar et al. \(2020\)](#) claim to have observed that nearly 100% of the non-social learners (those who did not post any comments), also did not complete any steps. If so, a valuable feature for prediction of attrition would be knowing whether the number of comments is zero. Another corollary of such observation is that an intervention based on textual analysis of the comments would be perhaps less valuable, especially when



considering that there is an important category of learners who do engage (and may complete) yet do not comment, as the converse is not necessarily true (i.e. it does not follow that all of those who do not complete, also did not post any comment).

The above has not deterred these researchers and many others in doing textual analysis of the comments, for prediction of attrition (c.f. [Duru, Sunar, White, and Diri \(2021\)](#) and others), and though there is research value on focusing solely on those who comment, for various valid, methodological reasons, I am interested in creating a more holistic model, in which all learners could be represented. This would encompass all they come to do in their courses, i.e. their interactions with the content matter too, not solely in their interactions with their peers. Particularly in the FutureLearn context, as [Sharples and Ferguson \(2019\)](#) note, “learning through conversation on MOOCs is not a replacement for direct instruction but an adjunct to it”.

Much of the recent literature looking into identification of sub-populations in MOOCs utilise social network analysis tools ([Gillani and Eynon, 2014](#); [Poquet et al., 2020](#); [Wise and Cui, 2018](#)), and though the identified groups are often comparable, I will not review this research in much detail as their methodology differs substantially from mine.

[Dowell et al. \(2018\)](#) considers conversations within a MOOC thread as possibly organized visually or temporally. In particular, for the visual organisation, the authors consider the order of the conversation as respecting the labelled dependency between posts and comments on posts, but ignoring the temporal order of the contributions. In their opinion, the temporal ordering more accurately represents the evolving development of learners’ ideas over time. Arguably, it does hide the true semantic relations as the context within which a comment is written (as part of the conversation) may be stripped or difficult to capture in the timeline view.

## 2.2.1 FutureLearn

FutureLearn is one of the top three global MOOC platforms, with fifteen million learners enrolled in 2020 ([Shah, 2020a](#)), only behind Coursera and edX in number of registrations. The University of Southampton is one of the founding partners of FutureLearn, joining the consortium in 2013, and with 22 courses offered to date, is still an important MOOC-providing institution on the platform.

### 2.2.1.1 Architecture

FutureLearn courses are organised in weeks. Each week contains a set of activities, called “steps”, each of which has a learning object belonging to a prescribed category. Typical examples of these categories are: videos, articles, exercises, discussions, reflections, quizzes and peer reviews. For each step, learners are able to write comments, each of these in turn can be visibly “liked” (as in mainstream social media platforms) and have replies or follow-up comments. This facility allows learners to build connections amongst the community and with the learning objects presented, as often these comments allow for their personal reflections and expressions of their own understanding (or lack thereof).

For each its courses, FutureLearn logs traces of learners activities, organising them into files, some of which consist of the following<sup>4</sup>:

- **Enrolments:** Entries of participants registered in the course containing demographic data.
- **Step Activity:** Most important course activity containing aggregated records for course step visits and completion actions.
- **Comments:** This file contains the social forum interactions classified per course, week and step.
- **Peer Review Assignments:** This file contains the assignment submission by the participant along with the relevant data to classify it correctly among the courses, weeks and steps.
- **Peer Review Reviews:** The assignment reviews are stored in this separate CSV file which also references the corresponding entry at the assignments file.
- **Question Response:** The quiz questions’ response attempts along with the outcome (correct/incorrect) are contained in this file.
- **Weekly Sentiment Survey Responses:** A experience rating provided by learners on specific weeks, with reasons for the feedback.

---

<sup>4</sup>This list might not be complete as designers in FutureLearn add functionality to the platform over time. Also, not every features are supported for all of their courses. For example, the MOOC “Understanding Language: Learning and Teaching” does not have Peer Review Assignments, and the MOOC “Archaeology of Portus: Exploring the Lost Arbour of Ancient Rome” had them in its early runs but ceased having them after the third run.

- **Video Stats:** A number of MOOC-level statistics for video types of steps including: duration, number of views, downloads, total caption views, total transcript views, device used, and views disaggregated by region and percentage of the video (with counts at 5%, 10%, 25%, 50%, 75%, 95% and 100%).
- **Archetype survey responses:** Learners responses to optional survey about their learning archetype, as described in Table 2.4 below.
- **Leaving Survey Responses:** Learners responses to optional leaving survey. This survey is offered to learners who leave a course, and it records the reason for leaving the course, the last completed step and week, and the timestamp for both the last step and the leaving event.

FutureLearn rolled out a survey in 2017 through all of their courses, and responses from nearly 7,000 learners were coded and organised in learner *archetypes*, which are defined by FutureLearn as “patterns of behaviour that others are likely to follow.” The driver behind such categorisation was to identify those learners who are more likely to bring revenue, so that they could focus their efforts in supporting these learners (Walker, 2018b). The archetypes identified, based on their motivations to pursue MOOCs, either for personal development employability and learning either to better understand (and “fix”) a problem or for the joy of learning itself, are the following seven: Advancers, Preparers, Explorers, Hobbyists, Vitalizers and Fixers. This survey is still being offered to participants joining a new course in FutureLearn to date, who are presented with a multiple-choice question, as per Figure 2.3, which signals the interest of the platform in categorising their learners. However, the answering rates to these surveys are considerably low, as Table D.3, shows. To give an example, for run 7 of the Understanding Language MOOC, there were only 606 survey participants, out of 6,033 enrolled learners. In fact, for all the courses where there is survey information collected, approximately 10% of enrolled participants take the survey, despite it being very simple, consisting of only one multiple-choice question, as shown in Figure 2.3, though this might be attributable to the funnel of participation effect, as coined by Clow (2013). Table 2.4 summarises the characteristics of these archetypes.

### 2.2.1.2 Design

The use of this architecture reflects the pedagogical underpinnings that informed FutureLearn’s design. It was based on an explicit pedagogy of conversational learning, resting on a framework by Laurillard (2013), who was in turn inspired by the conversational theory formulated by Pask (1975). The platform fulfills the requirements for this

## Tell us why you joined

What's your main reason for joining **Introduction to Virtual, Augmented and Mixed Reality?**

### I want to:

(Please choose one answer)

- Support my personal interests or community activities
- Prepare for a work or study goal, such as an interview or exam
- Improve my wellbeing
- Explore future work or study options
- Satisfy my curiosity and love of learning
- Understand or manage a situation in my personal life
- Develop or stay up to date in my field
- Other (please specify)

Knowing what motivates our learners helps FutureLearn understand you better and improve our courses.

For information about how your responses will be used and stored, please review our [Terms and conditions](#) and [Privacy policy](#).

FIGURE 2.3: Archetype survey question in a FutureLearn MOOC, as seen by the participant. Its answers, together with the demographics information collated at enrolment, help the platform characterise learners (Screenshot taken from a survey presented to myself on joining the course Introduction to Virtual, Augmented and Mixed Reality in December 2020.)

TABLE 2.4: Summary of the FutureLearn archetypes as researched by Tracy Walker and reported in the series of newspaper pages by Niam O'Grady (Walker, 2018a,b,c,d).

Archetype name	Comments and reported needs	Age groups	Jobs and education	Likely location	Certificate purchasing
Advancers	Need accreditation of work-related skills, with clear outcomes, on up-to-date information on trend topics, with pathways for specialisation.	26-35	full-time in employment	Asia (36%), Europe (30%), Africa (18%)	The most likely learners to purchase a certificate.
Preparers	As above, but also ways to build confidence in knowledge, such as tests. Support for non-native English speakers.	19-26	student or early career	Asia (47%)	Likely to purchase.
Explorers	As advancers, but with reassurance about the viability of their chosen path.	26-35	seeking career change	Asia (38%), Europe (34%), Africa (15%)	Less likely than the two above.
Hobbyists	Courses supporting their existing personal projects, leisure activities and pastimes. They had the best activation rate and the best full participation rate.	56-75	retired	Europe (61%)	Not reported.
Vitalizers	Lifelong learners, want to be indulged with stimulating topics. They had the highest number of enrolments of all archetypes.	56-75	retired	Europe (63%)	Not reported.
Fixers	To manage and understand personal issues (health, social, political or cultural). They want empathy, understanding, confidence and empowerment. Accessible content and expert advice.	all	varied	Asia (45%), Europe (28%)	Least likely to purchase a certificate.
Flourishers	As above, but more likely to take several courses, particularly on well-being, health and the arts. They want accessible content that can be consumed on the go.	all	varied	Europe (40%)	Not interested in certification.

framework, as it includes the following elements to support effective conversations for learning at scale (Sharples and Ferguson, 2019):

- a consistent language, articulated through a pattern library for the platform;
- it supports a variety of pedagogic elements, through different types of steps;
- supports conversations for action and description, in the way of comments and discussion steps;
- presents conversations in context, stimulated by its educators;
- which also encourage reflective conversations; share perspectives, via peer-review, discussion and even how feedback to quizzes is given;
- synthesising knowledge or reaching agreements;
- clear objectives and outcomes, allowing for learners to share their own goals;
- enabling educators to become facilitators of conversations; and finally,
- it has a structured content that can facilitate the tracking of conversations, via the inspection of the `comments.csv` file<sup>5</sup> provided in its datasets as well as a dashboard.

As explained before, with this approach, learning is the result of the social interaction between peers (and within the platform, the facilitators can act as such). Therefore, the platform has been built in order to afford this connectivist characteristic (and continues to be updated with new features that further support such affordances<sup>6</sup>).

---

<sup>5</sup>This and other files and their structure are presented in Figures 3.3 and 3.4.

<sup>6</sup>A recent innovation is the facility to work in small groups “to come together and reach shared understanding” (<https://www.futurelearn.com/about-futurelearn/our-principles>).

## 2.3 PeerWise

PeerWise is a web-based software that supports and manages the authoring and answering of multiple-choice questions (MCQs) for students within a cohort of a course. It is a widely adopted, free software, created and maintained by Paul Denny at the University of Auckland in New Zealand since 2008. PeerWise hosts over four million student-authored questions<sup>7</sup>, has received over ten million answers by students<sup>8</sup> and is used in more than 1500 institutions worldwide<sup>9</sup>, including many British institutions.

With regards to the disciplines it has been used for, it has been successfully used in various topics in Computer Science (e.g. Database development, Web systems development, Games AI, IT project management, as in [Devon, Paterson, Moffat, and McCrae \(2012\)](#)); in Health Sciences (e.g. Biochemistry ([McClean, 2015](#)), and Nursing education, as in [Rhodes \(2013\)](#) and [Rhodes et al. \(2015\)](#)) and other exact sciences (e.g. Physics and Organic Chemistry, [Mac Raighne, Casey, Howard, and Ryan \(2015\)](#)).

One of the insights from these reviewed works which is most relevant to my research is the observation by [Denny \(2013\)](#) that participation data in PeerWise is heavily skewed, with the most active 10% of students submitting approximately one third of all the answers. This suggests that even though PeerWise is a peer-supported digital environment designed to complement courses delivered face-to-face, learners engaging within this environment exhibit behaviours that are consistent to those exhibited by online learners, such as the 90-9-1 rule of online participation, discussed earlier (in subsection [2.1.3](#)).

The remainder of this section is organised as follows: Subsection [2.3.1](#) defines multiple-choice questions *MCQs*, as their production constitute the core learning activity supported by PeerWise. Subsection [2.3.2](#) discusses how can they be used for assessment in Computer Science education, thanks to the *affordances* of the software that are visible to the student, as well as those hidden ('under the hood') which can help study learning activity behaviours. In order to find reported research on learners behaviours whilst using PeerWise, and more specifically, gain insights on what features are used by the research community, I conducted a scoping review, and its methodology is described in Section [??](#). The research identified therein is summarised in Table [2.5](#).

---

<sup>7</sup>Denny, P. (July 2019) *PeerWise mini-symposium*. (<https://www.ucl.ac.uk/brain-sciences/events/2019/jul/peerwise-mini-symposium>, last accessed on 11 July 2020.)

<sup>8</sup>Denny, P. (n.d.) "If at first you don't succeed, answer again!" *PeerWise blog*. ([https://peerwise.cs.auckland.ac.nz/docs/community/if\\_at\\_first\\_you\\_dont\\_succeed/](https://peerwise.cs.auckland.ac.nz/docs/community/if_at_first_you_dont_succeed/), last accessed on 11 July 2020.)

<sup>9</sup>PeerWise. (<https://peerwise.cs.auckland.ac.nz/>.)

TABLE 2.5: Summary of relevant quantitative studies of PeerWise learners in the literature

Authors	Data description	Methodology	Features extracted or engineered from PeerWise data
Denny, Luxton-Reilly and Hamer (2008b)	Courses: CS111, CS105, CS220, EG131 (in Computer Science or Engineering, at Auckland) Weeks: 4-8 Learners: 469 in total.	Studied data from various courses where the use of PeerWise was compulsory, observing that most students contribute only the minimum requirement of questions but exceeded the requirement for answers. Commenting was not a compulsory activity (and was not marked) yet in all courses there were many. As the compulsory participation in all courses had deadlines, the period in which the activity was recorded was split in before and after the deadline for each course (teaching vs study time)	Learners are not categorised but their activity was recorded with the metrics: · Questions: Avg. 1.4-3.0 (teaching) vs 0-0.2 (study) · Answers: Avg. 19.8-43.0 vs 10.8-40.9 · Comments: Avg. 1.1-2.1 (with length correlated with exam marks by 34%).
Denny (2013)	Course: POPLHLTH111 (at Auckland) Weeks: 4 Learners: 1031, 516 exposed to a version of PeerWise with badges on and 515 without Participants surveyed: 519 (256 of which were from the <i>badges on</i> group).	Studied data from authentic student use of the tool, with half randomly presented its new version (with the possibility of earning badges) and the other half the older version. Statistics of the usage were compared amongst groups and a student perception survey was conducted after the data collection.	No categories of learners. Cohort activity was measured as: · Questions: 1,311 vs 1,306. Avg. 2.54 in each · Answers: 52,599 vs 43,086. Avg. 68 vs 60 · Days active: 52,599 vs 43,086. Avg. 7.01 vs 6.21 · Number of badges earned (80% earned 11 or more, 0.3% less than 7 and 0.1% all 22).
Biggins, Crowley, Bolat, Dupac and Dogan (2015)	Course: unspecified (at Bourmemouth) Weeks: 12 Learners: 50 out of 52 required to use PeerWise	Studied data from authentic student use of the tool, with half randomly presented its new version (with the possibility of earning badges) and the other half the older version. Statistics of the usage were compared amongst groups and a student perception survey was conducted after the data collection.	No categories of learners. Cohort activity was measured as: · Questions authored: 804 (Avg. 16.1) · Answers submitted: 3,345 (Avg. 66.9) · Answers per question: 3,273 (Avg. 4.1) · Questions ratings: 2,897 (Avg. 3.6) · Average ratings: 1.8 · Number of trophies (badges) earned: 941 (Avg. 18.8).
Doyle and Buckley (2020)	Courses: <i>Taxation</i> (at Limerick) Duration: One semester Learners: 240.	Quantitative. Students were randomly assigned a topic out of five to author questions about. They were encouraged to answer questions on all topics. Researchers found that students performed better in the exam on the topic areas they authored questions in.	No categories of learners. In addition to PeerWise features, the following metrics were calculated for each learner: · MarkNoMCQ (exam marks on the topics they did not author an MCQ in) · MarkWriteMCQ (remaining marks)



### 2.3.1 Multiple-Choice Questions

A multiple-choice question (MCQ) comprises of a statement or question (a stem), and its answer amongst a number of distractors. The student's task is to select the correct answer [Draper \(2009\)](#). Distractors of well-designed MCQs can expose different kinds of misunderstandings or common errors associated with the material being tested. For this reason, MCQs are widely used for summative and formative assessment, especially at early stages and in first- and second-year undergraduate studies. Typically, MCQs would be used to test factual knowledge (hence their prevalent use in STEMM, of which introduction to programming are prime examples ([Denny, Tempero, Garbett, and Petersen, 2017](#); [Luxton-Reilly, Denny, Plimmer, and Sheehan, 2012](#))), though they have also been used in more discursive disciplines ([Humpage et al., 2014](#); [Renzo et al., 2014](#)).

An advantage of MCQ-based assessment is that it allows for automated marking, and therefore instant provision of feedback to learners [Davies, Proops, and Carolan \(2020\)](#), as well as opportunities for multiple attempts, that is ideal for formative assessment, providing that there is opportunity for reflection on the mistakes made. However, the use of MCQs has been criticised on the grounds of fostering shallow or strategic learning [Biggs and Tang \(2007\)](#).

A way to overcome this problem, and fostering deeper learning, is the use of an environment in which the students do not simply answer the questions, but also create them and critique them, as within PeerWise, which can be beneficial to both ends of the ability spectrum in a course. For example, [Denny, Hamer, Luxton-Reilly, and Purchase \(2008a\)](#) studied learners' performance both within PeerWise and in the final exam (which had an MCQ component) in a first-year programming course, as well as their engagement in other learning activities (outside PeerWise, such as participation in labs and projects), concluding that it was both higher- and lower-achieving students who exhibited more learning gains through their interaction with PeerWise, whereas for mid-achieving students, their engagement in other learning activities may play a larger role. The conjecture they arrived at is that the higher-achieving students were more likely to benefit from the higher learning gains associated with engaging in question creation and critical appraisal of those made by fellow students, whilst the lower-achieving ones were more likely to improve their exam results through the learning by rota associated with the strategic learning from just answering MCQs in the question bank. For all kinds of students, however, there are huge benefits associated to the use of PeerWise as a peer-supported environment.

Though technically there is no mechanism within PeerWise for formal assessment of the questions therein produced, some higher education institutions assess (externally

to PeerWise) content created by students within the software, requiring them to self-select their own best contributions, and write a reflective essay about them, for example. However, because the engagement in these activities take place in the context of assessed courses, some authors have explored how students' engagement levels correspond with their grades, and PeerWise has been found to increase student engagement in course content (Denny, 2013), and improved exam results and positive reviews by students (Denny, 2013; Biggins et al., 2015; Renzo et al., 2014).

### 2.3.2 PeerWise affordances

Motivation for its use is supported by its many affordances. In particular, *Users* of the platform have the ability to gain *followers* and follow authors amongst their peers, which is an element present in many social media platforms. Another important affordance of this software is that, being generated by peers, the MCQs presented are critically appraised, rather than 'accepted as correct' as would typically be when created by the lecturer Luxton-Reilly et al. (2012). There are mechanisms to award *ratings* to questions by all users (where both the quality and difficulty of a given question are given a score), as well as to leave *comments* and *replies*.

### 2.3.3 Gamification elements in PeerWise

One of the main characteristics of PeerWise is that it uses gamification to nudge participation, which has proven empirically to have a positive effect on the number of *questions* students answer (Denny, 2013). Nudges for participation are given via notifications, which provide learners information about their own engagement, such as by the use of leaderboards, as well as via "badges". PeerWise uses social norms and salience by exposing to each student their relative level of engagement with the software by signalling their relative position within the cohort, through simple visualisations (see Figure 2.4). Badges can be acquired by students in reaching various milestones (some examples are in Figure 2.5)

#### Badges

To promote participation in PeerWise, special icons, called badges, are revealed to the student once they meet certain conditions. This is referred as "earning the badge". There are three kinds of badges that can be earned: *Basic*, *Standard* and *Elite*. *Basic* badges

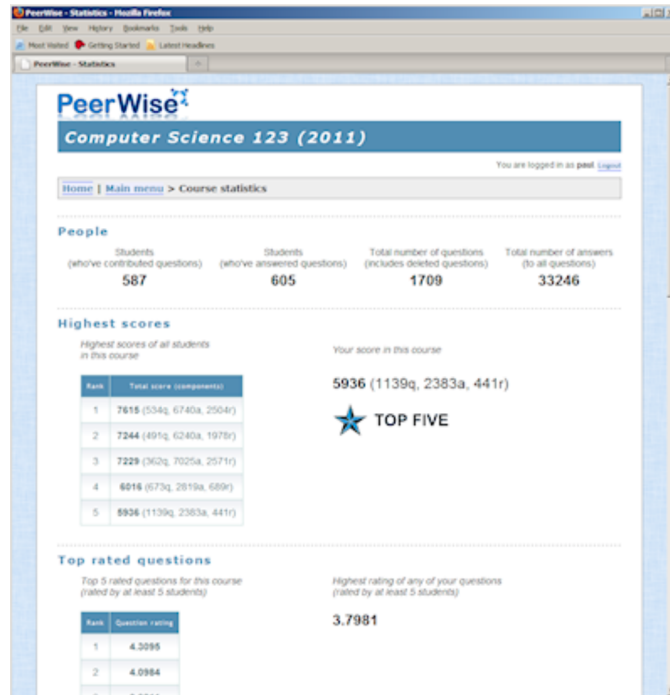


FIGURE 2.4: PeerWise leaderboard: Users can keep track of the top rated questions, the most popular contributors and the highest scores. (From <https://peerwise.cs.auckland.ac.nz/docs/students/>. Last accessed 2<sup>nd</sup> March 2021.)



FIGURE 2.5: PeerWise badges: Students gain badges for reaching different authoring, answering and commenting milestones. Lecturers can keep track of students' participation and engagement. (Images from <https://peerwise.cs.auckland.ac.nz/docs/students/>. Last accessed 2<sup>nd</sup> March 2021.)

typically mark the first time a particular interaction milestone has been achieved, meaning that they can only be earned once. *Standard badges* can be earned multiple times, meaning that the total number of badges can be greater than the number of different badges earned. Finally, *elite badges* are the most challenging to earn. Table 2.6 comprises all the badges that students can obtain through their engagement in the course through PeerWise, with a description of the ways these badges can be awarded to a student. At present, there are 25 possible badges to be earned, expanded from the original set of 22.

“Basic” badges are typically very easy to earn, with minimal engagement. These reward students as they explore the platform’s functionalities, in particular: authoring, answering and commenting on questions, agreeing on (and replying to) comments written by others, follow others and verifying others’ answers. Denny (2013) reports 80% of 516 learners that were able to earn badges in PeerWise (the “*badges on group*”) earned more than 11, with only one student earning them all and three earning less than the seven basic ones<sup>10</sup>.

## 2.4 Other peer-supported digital environments for face-to-face instruction

As seen in the literature review covered in Section 2.3, I did not find any prior research on classification of PeerWise learners using clustering, which is the approach I use in this thesis. Indeed, there are challenges in securing educational datasets on face-to-face instruction contexts, as indicated by Sarker (2014). The heterogeneity of technologies used for capturing traces of relevant learning activity, together with the institutional challenges associated with the ethical use of such traces (Slade and Prinsloo, 2013) result in the relatively scarce availability of well-structured, coherent datasets for research. Those that exist, tend to be around the use of Learning Management Systems (LMS), course managements systems (CMS) or virtual learning environments (VLE) supporting face-to-face instruction. These are many systems used in the context of educational institutions offering technology-enhanced learning or computer-assisted instruction – Blackboard™, Canvas, Moodle and Sakai are well-known examples, though there are others, such as SkyPrep, and Docebo (both commercial products) and open-source technologies, developed in-house by higher education institutions and made publicly available such as ILIAS and Online Learning and Training (OLAT). Godwin-Jones (2012)

---

<sup>10</sup>At the time of the study by Denny (2013), only 22 badges out of the present 25 were available, seven of which were basic ones.

TABLE 2.6: Badges currently available in PeerWise for a student  $s$ . The column *Type* identifies the kind of interaction a badge rewards: question authoring (Q), question answering (Q), comments (C), replies (R), and consistency over time or accuracy ( $\dagger$ ). A ‘yes’ in the column *New* indicates that the badge was not present in the original research paper by [Denny \(2013\)](#)

Category	Badge name	Description	Type	New
Basic	“Question author”	$s$ contributed one question	Q	
	“Question answerer”	$s$ answered one question	A	
	“Star-crossed”	$s$ agreed or disagreed with a comment	C	
	“Comment”	$s$ wrote one comment	C	
	“Author-reply”	$s$ replied to a comment written about own question	R	
	“Follower”	$s$ followed one or more authors		
	“I’ll be back”	$s$ answered correctly ten or more questions, on each of three different days)	A $\dagger$	
	“Verifier”	$s$ has confirmed one answer or more		yes
Standard	“Helper”	$s$ improved the explanation of an existing question	C	
	“Popular question author”	per question authored by $s$ that was answered ten times or more	Q	
	“Discussed question author”	per question authored by $s$ that received two or more comments	Q	
	“Commentator”	$s$ wrote five comments or more		
	“Critique”	$s$ agreed or disagreed with ten comments	C	
	“Rater”	$s$ submitted a rating for ten questions		
	“Scholar”	$s$ answered ten questions correctly	A	
	“Commitment”	$s$ answered correctly ten or more questions, on each of five consecutive days	A $\dagger$	
Elite	“Good question author”	per question authored by $s$ rated as excellent five times or more	Q	
	“Insight”	$s$ wrote two or more comments that someone agreed with	C	
	“Conversation”	$s$ replied to five comments about own questions	R	
	“Genius”	$s$ answered ten questions in a row correctly	A $\dagger$	
	“Leader”	$s$ had one or more followers as a question author	Q	
	“Einstein”	$s$ answered 20 questions in a row correctly	A $\dagger$	
	“Obsessed”	$s$ answered correctly ten or more questions, on each of ten consecutive days	A $\dagger\dagger$	
	“Super scholar”	$s$ answered correctly a total of 50 questions	A	yes
	“Legend”	$s$ submitted a correct answer on 31 distinct days	A $\dagger\dagger$	yes

provide a comprehensive (if somewhat dated) list and critique of a number of such systems.

In terms of research on data generated in such environments, [Romero and Ventura \(2010\)](#) reviewed 304 studies indicating that students use LMS to personalise their learning, reviewing specific material and engaging in relevant discussions as they prepare for their exams. Lecturers and instructors use them to give and receive prompt feedback about their instruction, as well as to provide timely support to students (e.g. struggling students need additional attention to complete their courses more successfully ([Baepler and Murdoch, 2010](#)), as the failure to do so comes at a great cost, not only to these students but to their institutions). Administrators use LMS to inform their allocation of institutional resources, and other decision-making processes ([Romero and Ventura, 2010](#)). These authors argue the need for the integration of educational data mining tools into the e-learning environment, which can be achieved via LMS.

One rare example using clustering on data from face-to-face instruction courses is that by [Bogarín, Romero, Cerezo, and Sánchez-Santillán \(2014\)](#). They report use of

clustering for improving modelling for process mining in an online course using the VLE Moodle 2.0 to complement a face-to-face undergraduate course in Psychology. In their study, they used the Expectation Maximisation clustering algorithm from WEKA to find three clusters of distinguishable behaviours. One of these clusters contained the majority of the failing students in the course, whilst the majority of the passing students were in the other two. In the first case, students models of interactions in Moodle were simpler and more limited, whereas those in the other two clusters were much more complex. The first of the “passing” clusters grouped learners with more strategic actions within the VLE than the second cluster, so there were observable differences of behaviour even with similar levels of attainment. This finding suggests, not only that the patterns of engagement with the digital environment can be used to predict attainment, but that clustering algorithms are a useful tool to identify nuanced behaviours.

There is one final aspect regarding peer-supported digital environments, that is of relevance to this thesis. Regardless of their implementation details, these environments include mechanisms such as discussion forums within which learners can communicate. Their organisation details vary but it is very common that they are supported by multi-level structure, such as thread, posts and replies, which are featured not only in VLEs like Blackboard and others but also in MOOCs like Coursera and edX. Though the structure itself can be a useful indicator of the conversational patterns of interaction, much research is focused on analysing unstructured learner posts, as seen in the next section.

## **2.5 Learning analytics, educational data mining and academic analytics**

Modern digital technologies are characterised by a high integration of information processing, connectivity, media storage and even sensing capabilities, making it easier than ever to collect, analyse and exchange data about our daily activities. For higher education students, this means that they also generate a rich data trail as they navigate their way through towards successful completion of their studies, particularly in the contexts where these digital technologies have been adopted. In more recent times, such adoption has been on the increase as institutions are driven by the need to facilitate the delivery of learning content as well as the assessment of students’ work, often remotely even in the contexts of face-to-face courses, giving further opportunities for learners to add to the already rich data trail of activity aforementioned as they measure engagement, attendance and attainment of learning.

Learning analytics, educational data mining and academic analytics are all, in the

broadest terms, similarly concerned with the analysis of student records held by the institution, as well as course management system audits, including statistics on online participation and similar metrics. They do so to either understand learners and facilitate interventions (in the case of learning analytics), or to inform stakeholders decisions in HE institutions (in academic analytics). Educational data mining originally differentiated from the other two mainly through its methods, which were borne out of the artificial intelligence and data mining communities. In more recent times the divisions amongst these disciplines are increasingly blurred and arguably they are more related to academic communities membership rather than marked differences in methodologies or utility of the findings.

In particular, Learning Analytics was first defined in the call for papers for the first Learning Analytics and Knowledge conference, in Banff, Canada, by the Society for Learning Analytics Research (SoLAR), as: “... *the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs*” (as cited in [Siemens and Baker \(2012\)](#)). As such, it is a multi-disciplinary field, concerning often the expertise of data scientists, learning technologists, psychologists, educators, educational domain content experts, computer scientists and even measurement specialists (as illustrated in [Figure 2.6](#)).

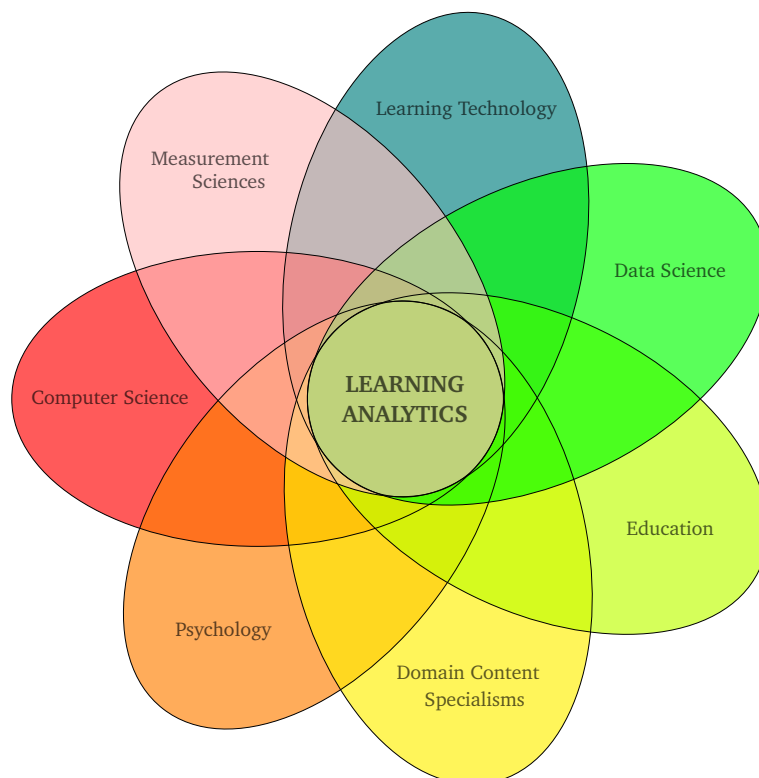


FIGURE 2.6: Learning Analytics is a multidisciplinary field

Educational data mining involves processing such data (collected from the VLE or other sources) through machine learning algorithms, enabling knowledge discovery, which is “the non-trivial extraction of implicit, previously unknown, and potentially useful information from data” (Frawley, Piatetsky-Shapiro, and Matheus, 1992).

Academic analytics are similarly considered as useful tools to study scholarly innovations in teaching and learning (Baepler and Murdoch, 2010), but the level or object of the analysis tends to be at least of institutional scale (Sharkey, 2011), but often, regional, national and even international scale. According to Baepler and Murdoch (2010), the term *academic analytics* was originally coined by the makers of the virtual learning environment (VLE) Blackboard™, and it has become widely accepted to describe the actions “that can be taken with real-time data reporting and with predictive modeling” which in turn helps to suggest likely outcomes from certain behavioural patterns.

The kinds of problems studied in these fields include retention, attrition (Glynn, Sauer, and Miller, 2003), and dropout (e.g. Barber and Sharkey (2012)). Through the understanding of learners’ progress and engagement using these techniques, it becomes possible to plan and deliver personalized interventions: be it directly, in the form of “nudges”<sup>11</sup>, or indirectly, via institutional processes, offering the opportunity to promptly identify performance issues so that actions can be taken to encourage student success. This is indeed the ultimate goal of learning analytics, that it provides actionable insights that can assist educators in supporting learners, as well as informing policies and reforms that can effect change to the benefit of the many. As Siemens and Long (2011) put it, by “penetrating the fog” that is over much of higher education about the mechanics of supported learning, foster a better understanding of how learners learn.

One of the challenges in learning analytics as it transitions into a rigorous, empirical discipline, is the creation of frameworks that facilitate replication work in learning analytics, which it is still very unusual (Dowell and Poquet, 2021).

### 2.5.1 Feature Engineering

Much of educational analytics research (as well as data mining research) rests on the goodness of features from the data, chosen for the application of methods within these fields. The goal of feature engineering in learning analytics is a “greater interpretability, generalizability, transferability, applicability and with clearer evidence for effectiveness”

---

<sup>11</sup>As mentioned in the introduction, I explored in a position paper the idea of using learning analytics for behavioural change to address attrition in MOOCs using persuasive technology (Wilde, 2016).



(Baker, 2020). The design of predictor variables is surprisingly poorly studied and documented (Baker, 2020). It is said to involve lore rather than well-known and validated principles. A well-designed feature, however, is one that is meaningfully linked to the construct under study and is potentially interpretable by humans. Baker (2020) discourages feature engineers from engaging in deep learning models that are not interpretable and may have problems with overfitting and smaller datasets than the algorithms assumptions lie on. Similarly, though much research includes them, Baker considers the use of “tautologies” from the data as poor feature-engineering design, such as, for example, the final course grade from assignments and tests. Good heuristics for the design of features are offered as follows:

1. Brainstorm with domain experts: deferring judgement, encouraging lateral thinking but building on the ideas of other researchers. As many features as possible from the data that can be collected.
2. Decide what features to create from the brainstorm, biasing in favour of tractability and diversity of features.
3. Create the features, preferably with a scripting language for reproducibility.
4. Study the impact of features on model goodness, possibly with a confusion matrix. Also decide what to do with outliers (typically instances above or below three standard deviations from the mean), ideally visually, such as through the use of box-and-whisker plots.
5. Iterate on features if useful, splitting the data into groups of interest, and modelling each subgroup.
6. Go back to step 3 (or even 1).

Baker (2020) also notes that it is possible to become a domain expert, by understanding the literature, having conducted classroom observations, or having had teaching experience relevant to the construct under study. However, conversations with others are helpful.

An alternative way to produce features is through a process Baker calls “knowledge engineering” where a construct is modelled with a domain expert, features are chosen and recombined within machine learning. Though more elaborate, this process can lead to better performance in unseen data or in data across different sources than a more traditionally-developed feature set (Baker, 2020; Paquette and Baker, 2019).

## 2.5.2 Clustering

There are several algorithms used in the literature, as surveyed by [Dutt, Aghabozrgi, Ismail, and Mahrooian \(2015\)](#) in the context of educational data mining. These typically fall in one of the two types: *partitioning* and *hierarchical*. Partitioning clustering algorithms part from the principle that there is a number of clusters that the data is known to fall into (e.g. the *k-Means* clustering algorithm, and its variations, such as *X-Means* clustering, by [Pelleg and Moore \(2000\)](#)).

This requires both a good knowledge of the domain where the data is drawn from and a relative noise-free dataset. Another assumption, typically an implementation limitation, is that the size of the dataset is sufficiently small to wholly reside in memory because of the way the assignation to clusters is typically done. Hierarchical algorithms, on the other hand, presume the assignation to clusters to follow some kind of hierarchy, and though the number of clusters is not required to be known a priori, a stopping condition might have to be defined, particularly in the case of *divisive* hierarchical clustering. In this case, the dataset is separated in successive steps depending on a measure of distance, upon which the stopping condition can be defined. Therefore, this is a top-down approach.

The choice between clustering algorithms is ultimately dependent on the problem and the domain, as there is no single clusterer that outperforms all others in all situations. Hence, performance metrics must be applied to conduct a comparison and make a selection.


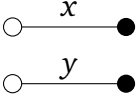
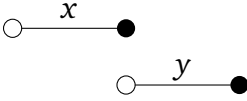
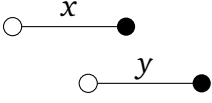
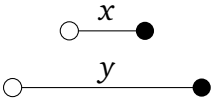
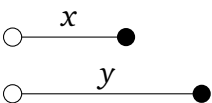
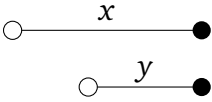
In the context of learning analytics, [Baker \(2020\)](#) suggests that the ultimate arbiter on the goodness of a clustering algorithm is whether the resulting clustering are interesting, rather than the actual fitness of the data to clusters. Poor interpretability of models is in fact a known limitation of the application of many clustering techniques, as observed by [Liu and Koedinger \(2017\)](#), “tend to result in student groupings that are difficult to interpret”. Yet for learning analytics, interpretation is key, especially if the observations inform policy. This is an observation to keep in mind when choosing to apply a clustering algorithms on educational data.

## 2.6 Interval Algebra

A well established interval-based temporal logic is commonly known as “Allen’s interval algebra” ([Allen, 1983](#)). hailed for its expressive power and simplicity, which facilitates

automated reasoning in applications over many domains where temporal hierarchies can be defined (Janhunen and Sioutis, 2020).

TABLE 2.7: Interval algebra: the thirteen possible relations (adapted from Allen (1983) and Hunsdale et al. (2017)). All relations have an inverse, except for “is equal to”. Not listed under “Endpoint conditions” are  $x_{\circ} < x_{\bullet}$  and  $y_{\circ} < y_{\bullet}$ . These apply to all.

Relation	Symbol	Inverse	Pictorial example	Endpoint conditions
$x$ precedes $y$	$<$	$>$		$x_{\bullet} < y_{\circ}$
$x$ is equal to $y$	$=$			$x_{\circ} = x_{\bullet} \wedge y_{\circ} = y_{\bullet}$
$x$ meets $y$	$m$	$mi$		$x_{\bullet} = y_{\circ}$
$x$ overlaps $y$	$\circ$	$oi$		$x_{\circ} < y_{\circ} < x_{\bullet} < y_{\bullet}$
$x$ is during $y$	$d$	$di$		$y_{\circ} < x_{\circ} < x_{\bullet} < y_{\bullet}$
$x$ starts $y$	$s$	$si$		$x_{\circ} = y_{\circ}$
$x$ finishes $y$	$f$	$fi$		$x_{\bullet} = y_{\bullet}$

The domain  $D$  of interval algebra is defined to be the set of intervals on the line of rational numbers, i.e.,  $D = \{x = (x_{\circ}, x_{\bullet}) \in Q \times Q \ni x_{\circ} < x_{\bullet}\}$ . Each base relation can be defined by appropriately constraining the endpoints of the two intervals at hand, which yields a total of 13 base relations comprising the set  $B$  defined as:  $B = \{e, p, pi, m, mi, \circ, oi, s, si, d, di, f, fi\}$ ; where  $p$  = precedes,  $e$  = equals,  $m$  = meets,  $\circ$  = overlaps,  $d$  = during,  $s$  = starts, and  $f$  = finishes respectively, with  $oi$  denoting the converse of  $\circ$  (note that  $ei = e$ ). For example,  $d$  is defined as  $d = (x, y) \in D \times D \mid x_{\circ} > y_{\circ} \wedge x_{\bullet} < y_{\bullet}$ ;  $x_1 < x_2$  ( $x_1$  takes place before  $x_2$ );  $x_1 \circ x_2$  ( $x_1$  overlaps  $x_2$ ); amongst others.

The relevance of these definitions will become evident when I introduce the model

of learner engagement within peer-supported learner environments in Chapter 4.

## 2.7 Information retrieval terminology

The sensitivity (or *recall*) in information retrieval is the rate of retrieved elements that are relevant amongst the total number of relevant elements [Witten, Frank, Hall, and Pal \(2017\)](#). Its counterpart definition in classification with machine learning is the proportion of elements of a given class that are correctly classified. A perfect recall therefore will have a value of one, and it will occur when there are no false negatives, as per the Equation 2.1. It means that all relevant elements were retrieved (and there are no false negatives).

$$Rec = \frac{TP}{TP + FN} \quad (2.1)$$

The precision is the rate of retrieved elements that are relevant amongst the total number of elements retrieved, as per the Equation 2.2. It means that all retrieved elements were relevant (and there are no false positives).

$$Prec = \frac{TN}{TN + FP} \quad (2.2)$$

The accuracy *Acc* (Equation 2.3), the *F1* score (Equation 2.4), and the Area Under the Receiver Operating Characteristics Curve (AUROC, or ROC Area) are provided in *scikit-learn* and WEKA. The ROC Area is used to evaluate classifiers' performance which is used in pattern recognition and machine learning ([Fawcett, 2006](#)). In simple terms, a ROC Area close to the value of one is indicative of a well-performing algorithm, with true-positive and true-negative rates consistently high.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.3)$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (2.4)$$

### 2.7.1 Multi-Class Approach Considerations

In order to evaluate the performance of such classification, it is possible to use metrics such as Recall, Precision, F1-score and AUROC. However these are defined for two-class

classification problems and its extension for multi-class problems could be done using the calculation of the *macro* or the *micro* score for Recall, Precision, F1-score and AU-ROC. This is the average metric per class which gives the same importance for each class. The other solution is the *micro* score which average the metric by giving more importance to the amount of data per class. In a multi-class problem, the Recall for a specific class is calculated against all the others together as if they were one class. Matches for this specific class represent the positive cases and matches for the combined class represent the negative cases. Applying this step for each class offers four different Recalls, which then are averaged using the previously explained macro score, as per Equation 2.5. A similar process is applied for Precision and F1-score, as shown in Equations 2.6 and 2.7.

$$Rec_{macro} = \frac{1}{|Class|} \times \sum_{i=1}^{|Class|} \frac{TP_i}{TP_i + FN_i} \quad (2.5)$$

$$Prec_{macro} = \frac{1}{|Class|} \times \sum_{i=1}^{|Class|} \frac{TN_i}{TN_i + FP_i} \quad (2.6)$$

$$F1-score_{macro} = \frac{1}{|Class|} \times \sum_{i=1}^{|Class|} \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i} \quad (2.7)$$

## 2.8 Summary

I reflect on the research questions listed in Section 1.2 to inform my literature review, as follows:

**RQ1:** *How can learner engagement be meaningfully compared across peer-supported digital environments?*

The literature reviewed around engagement and philosophies of learning presented in Section 2.1 discuss aspects of these environments that are essentially platform-agnostic. This means there is already a common understanding of the concerns around the phenomena occurring within peer-supported digital environments that can be used in such comparisons, such as the nature of the interactions occurring within, irrespective of the implementation details. However the review has also highlighted the challenges around the proliferation of these types of environments, such as poor interoperability and others which arise from the diversity of implementations. These call for a formal model of interactions, and Section 2.6 on Allen's algebra provided the language through which some of the formalisms required for a model of learner interactions can be articulated.

Also, the dialogic analysis for engagement presented by [Chua et al. \(2017\)](#), though developed on data from a FutureLearn MOOC, can be part of a model that can be extended and generalised to other platforms.

In particular, where both the dialogic and the temporal aspects of learning activity in theory can be used within an operationalised view of engagement. That operationalisation can be expressed and applied, in turn, to any practical examples in real-world platforms, irrespective on how the capture of the timestamps on the digital traces of activity is performed, or indeed, of any other implementation details. This forms the bases of a model that is presented in Chapter 4.

**RQ2:** *What does a data-driven approach to learner interactions reveal about learning engagement within FutureLearn MOOCs?*

Table 2.3 presents several ways in which researchers categorise MOOC learners and their engagement based on their interactions in the platform. Amongst those categories found, many relate to the level of engagement as well as the type of engagement and the time within which these interactions occur. Of particular interest and relevance to this thesis is the dialogic categorisation of learners by [Chua et al. \(2017\)](#). In terms of the methodology to categorise learners, clustering is an effective approach, in particular, k-Means clustering, as used by [Kizilcec et al. \(2013\)](#); [Ferguson and Clow \(2015a\)](#); [Tseng et al. \(2016\)](#); [Dowell et al. \(2018\)](#), but also expectation maximisation (EM), as used by [Bogarín et al. \(2014\)](#) and others. The number of clusters range between three ([Dowell et al., 2018](#)) and ten ([Ferguson and Clow, 2015a](#)), and typically dominated by dropouts across many of the studies in Table 2.3. A couple of the surveyed studies looked into orthogonal classifications ([Anderson et al., 2014](#); [Sunar et al., 2020](#)) but the majority looked at analysing all of the feature space and have only one classification, based on frequency or types of interactions.

**RQ3:** *What does a data-driven approach to learner interactions reveal about learning engagement within the PeerWise digital environment for face-to-face instruction?*

Table 2.5 lists features used in the literature for analysing student engagement within this platform. Amongst those features, the counts of questions, answers and comments by students is found as informative. Other features that characterise engagement in this platform are the badges earned (from those in Table 2.6) and the cohort activity.

**RQ4:** *Is learner engagement different in different kinds of peer-supported digital environments, be it a complement to face-to-face instruction, or a fully online course?*

Past research on learning analytics (Section 2.5) has shown how engagement is often studied, but a comparison between such different environments did not emerge in

the scoping reviews conducted. Unsupervised learning can elucidate this answer, in particular clustering, which was presented in Section 2.5.2.

There is value in using datasets from diverse platforms and disciplines to test the effectiveness of the operationalised metrics and hence the validity of a platform-agnostic model. As Duru et al. (2021) have also observed in their own literature survey, the research community generally apply their algorithms on a very limited number of courses, which can be observed also in most of the literature listed both in Tables 2.3 and 2.5. As a consequence, the findings are often not generalisable, with many threats to validity as such studies can be very platform-specific, discipline-specific, or even cohort-specific.

My research seeks to overcome such threats by not only using data from various offerings of the same course on the same platform, but looking at data from at least two distinct platforms, from at least two distinct disciplines and at least two consecutive offerings of each course. This approach was adopted so that we might also investigate the generalisability of the model as well as its accuracy. Further, I also sought to replicate and extend reported work by Chua et al. (2017). This is in itself is an interesting contribution of particular challenge, as mentioned before, replication work in learning analytics is still regarded as unusual (Dowell and Poquet, 2021). The way this is done here is described in general in Chapter 3 and in detail in Chapters 5 and 6.





# Chapter 3

## Methodological framework

*“If we can really understand the problem,  
the answer will come out of it,  
because the answer is not separate from the problem.”*

Jiddu Krishnamurti (b.11 May 1895–d.17 February 1986), quoted in “HOW TO SOLVE IT: MODERN HEURISTICS”, by Z. Michalewicz and D. B. Fogel, Springer.

In this thesis I have investigated learner engagement in peer-supported digital environments used in diverse contexts. In particular, within massive open online courses (MOOCs) in the context of purely online learning, and within PeerWise as an online environment embedded in a taught course at a higher education institution where the predominant modality of interaction is face-to-face instruction.

This chapter offers details of the methodology used in the development of this thesis, starting with a high-level description of its approach, in Section 3.1, both from a motivational and an operational perspective. Then, in Section 3.2, I offer details of the approach, through a mapping to a data-science pipeline. Finally, a summary of the methodology is presented in Section 3.3.

### 3.1 High-level description of the approach

The overarching theme amongst the literature surveyed in Sections 2.2 and 2.3 in the previous chapter was on learner engagement in peer-supported digital environments such as PeerWise and MOOCs (with a focus on FutureLearn MOOCs in particular),

though these are manifested differently in each context, partly due to the characteristics of the populations of learners but also due to differences in applied pedagogy and platform user-interface design. Differences aside, both FutureLearn and PeerWise capture data related to learners' participation in their courses. These data traces, left behind by participating learners, may allow institutions to assess learning gains, but also, to understand whether there are any sub-populations of interest worth identifying in order to inform both the course design and also possible interventions to increase engagement and facilitate learning.

In general terms, as proposed in the introduction of this thesis, the main aim of my doctoral research is the formulation of a model for analysis of learner activities in peer-supported digital environments. The purpose of such a model is to facilitate discussion regarding patterns of engagement irrespective of the platform within which it takes place. Crucially, such discussions could contribute towards discerning the parallels and contrasts between the behaviours exhibited by students in formal contexts (that include online activities even if it may primarily follow a face-to-face instruction model) and MOOCs, which are purely online, and may be part of non-formal or informal learning, as illustrated in Figure 3.1.

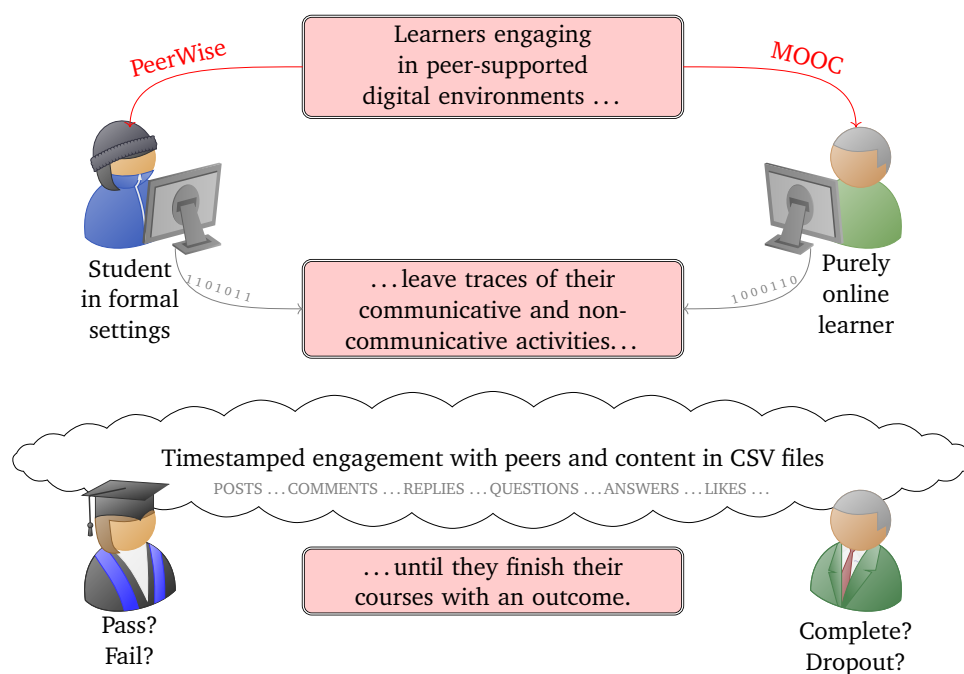


FIGURE 3.1: High-level view of scenarios spanning two different educational contexts. Both show learners within peer-supported digital environments, leaving a data trail of their engagement in the form of timestamped records of activity (e.g. when creating content or interacting with their peers through comments and replies). These traces, alongside the learners' outcomes, are captured in various CSV files in the respective environments (attainment or completion) for as long as they are in their courses.

In practical terms, the platform data is collated by both MOOC providers and given to the subscribing institutions with a structure that specifically affords the study of behavioural characteristics of the learners in the course (e.g. their graded achievements or the social interactions of the learners). This wealth of data offers great opportunities for collecting learning analytics, however, there are challenges in aligning the data collected and performing comparison studies not only because of the fundamentally different approaches taken in each case but also important differences in the technical implementations adopted.

In principle, a common approach is viable, since both contexts present learners engaging in a peer-supported digital environment, leaving a data trail of their engagement in the form of timestamped activity (creating content and interacting with their peers through comments and replies, for example). These digital traces are captured in various CSV files in the respective systems, which in turn can be analysed against the learner outcomes as illustrated in Figure 3.1. However, in practice, there are many challenges to overcome, related to data acquisition, processing, and representation in these environments, as well as in differences in the semantics of the data being captured, meaning that it is necessary to apply a data science approach.

## 3.2 A quantitative approach for data science

This thesis uses quantitative research methods, following the data science pipeline shown in Figure 3.2, and described in detail in subsections 3.2.1 to 3.2.5.



FIGURE 3.2: Data science pipeline applied in this study: experimental setup, data collection, data cleaning, feature extraction, feature selection, classification/clustering, analysis, evaluation of results, insights

### 3.2.1 Ask question

As per the epigraph in this Chapter, an important first step in the research is the understanding of the problem at hand, given that the answer is not separate from it. The research questions RQ1-RQ4, listed in Section 1.2 in the Introduction, are in fact aspects of the overarching question driving all research (“what do you want to know?”). All of

my research questions have at heart learner engagement on peer-supported digital environments, and a categorisation of such learner engagement in diverse environments.

In order to formulate the model (as detailed in Chapter 4), I first needed to understand how engagement in digital environments can be operationalised both in formal instruction and in MOOCs. This was achieved through the literature review presented in Chapter 2, with attention to the data characteristics and features used by others in the research community.

### 3.2.2 Collect data

In addition to the data collected via the scoping reviews mentioned above, for this thesis it was necessary to collect data from courses in the peer-supported environments of interest. Following an opportunistic approach to data collection, I was able to request MOOC data from FutureLearn courses provided by my institution, as well as that from students in one of my face-to-face courses. I was firstly required to submit an ethics application form for secondary data analysis. Though the details can be seen in Appendix A, it is important to describe here the engagement data I collected or requested access to. Following a rigorous process, once the ethics application was approved, a Data Protection Impact Assessment was conducted (see Appendix B) and a detailed Data Management Plan was provided (see Appendix C).

This process allowed me to study participation and attainment data for a total of 271,851 learners from nineteen courses provided by the University of Southampton between 2014 and 2019. The sampled courses encompass topics on human-computer interaction, archaeology, and pedagogy of language teaching, with seventeen of these being massive open online courses (MOOCs), and two in a face-to-face setting that included activities within a peer-supported digital environment. The data collection processes for each are described in the following subsections.

#### Collecting MOOC data

The first seventeen comprised two distinct FutureLearn courses, with both of them having been run multiple times over the considered period (eleven and six times respectively). At the time of the study, each MOOC run enrolled several thousands of learners from around the world (ranging between 1,286 to 58,782 across the seventeen runs), whereas the face-to-face courses enrolled 320 (140 and 180 respectively), also with an

international demographic, including students from Bulgaria, India, China and several other countries though primarily from the UK.

Figures 3.3 and 3.4 show the associated files for the FutureLearn MOOCs datasets, though it is worth to point out that FutureLearn does not collect all of the listed data files for all courses, nor is it available for every run of courses, as new affordances to the platform are progressively incorporated over time, and providers adapt the learning design for their existing courses. More details about these files can be found in Appendix D.

### Collecting PeerWise data

The second part of the engagement data concerned students enrolled on consecutive offerings of a module in a formal setting, following primarily a face-to-face instruction model which was complemented with the use of PeerWise as a peer-supported digital environment for creating, answering and critiquing multiple-choice questions. These modules ran in the academic years of 2015/16 and 2016/17 at the University of Southampton, and they are regarded as an opportunistic sample, given that I designed the assessments and delivered both as part of a team. Still, I sought permission for the ethical use of this data as appropriate and this was granted by the institution as detailed in Appendix A.

More specifically, these data subjects were enrolled on the second-year module “Interaction Design (COMP2213)”, a compulsory module for Computer Science. There are two data sources for this part of the study: firstly, students’ participation in the module via the free software PeerWise, which on registration students agreed it could be used for research purposes (previously approved as ERGO/FPSE/20318). Secondly, their attainment data which have been used to evaluate their learning within the module. The data in PeerWise are predominantly quantitative, reflecting their engagement with the module by their timestamped activity (e.g. creation of multiple-choice questions, provision of answers, ratings given and received on created questions, comments, number of replies given, number of followers, badges obtained, and so on). To these, as mentioned, the marks obtained in all elements of the assessment were also used. Figures 3.5 and 3.6 show the associated files for the PeerWise dataset (and assessment data).

The samples included timestamped digital traces of activity and comments generated by learners who completed at least one learning step (in the case of a MOOC learner) or activity, such as logging into the platform at least once (in the case of Peerwise). Achievement data for a learner consisted of what percentage of the course’s steps were completed by the learner in the case of MOOCs (with 50% being the minimum required

FIGURE 3.3: Associated Files for the FutureLearn MOOCs datasets (part I)



FIGURE 3.4: Associated Files for the FutureLearn MOOCs datasets (part II). Generated by FL only for run 6 of Portus and runs 8 and 9 of Understanding Language)

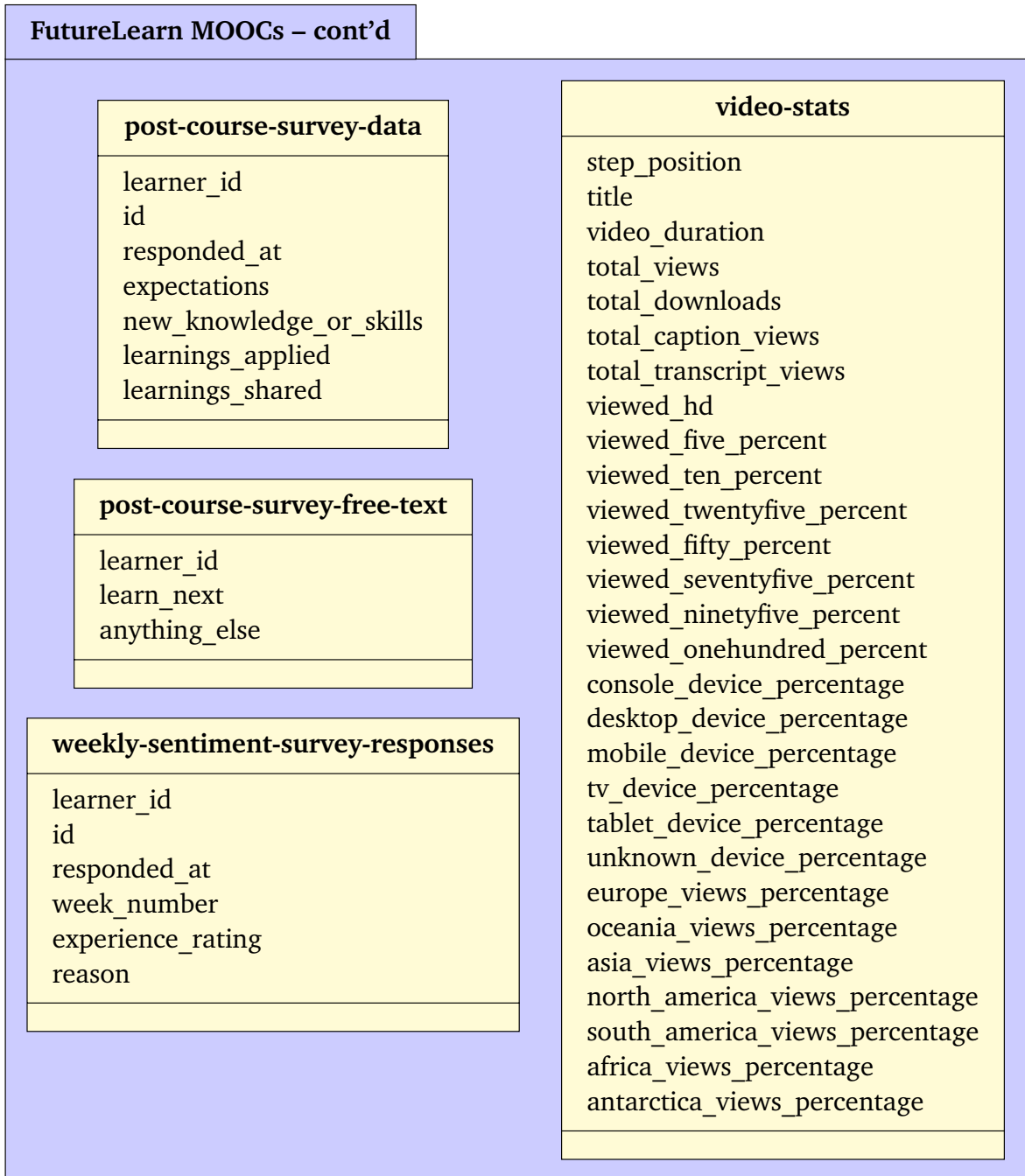


FIGURE 3.5: Associated Files for the PeerWise datasets (I)

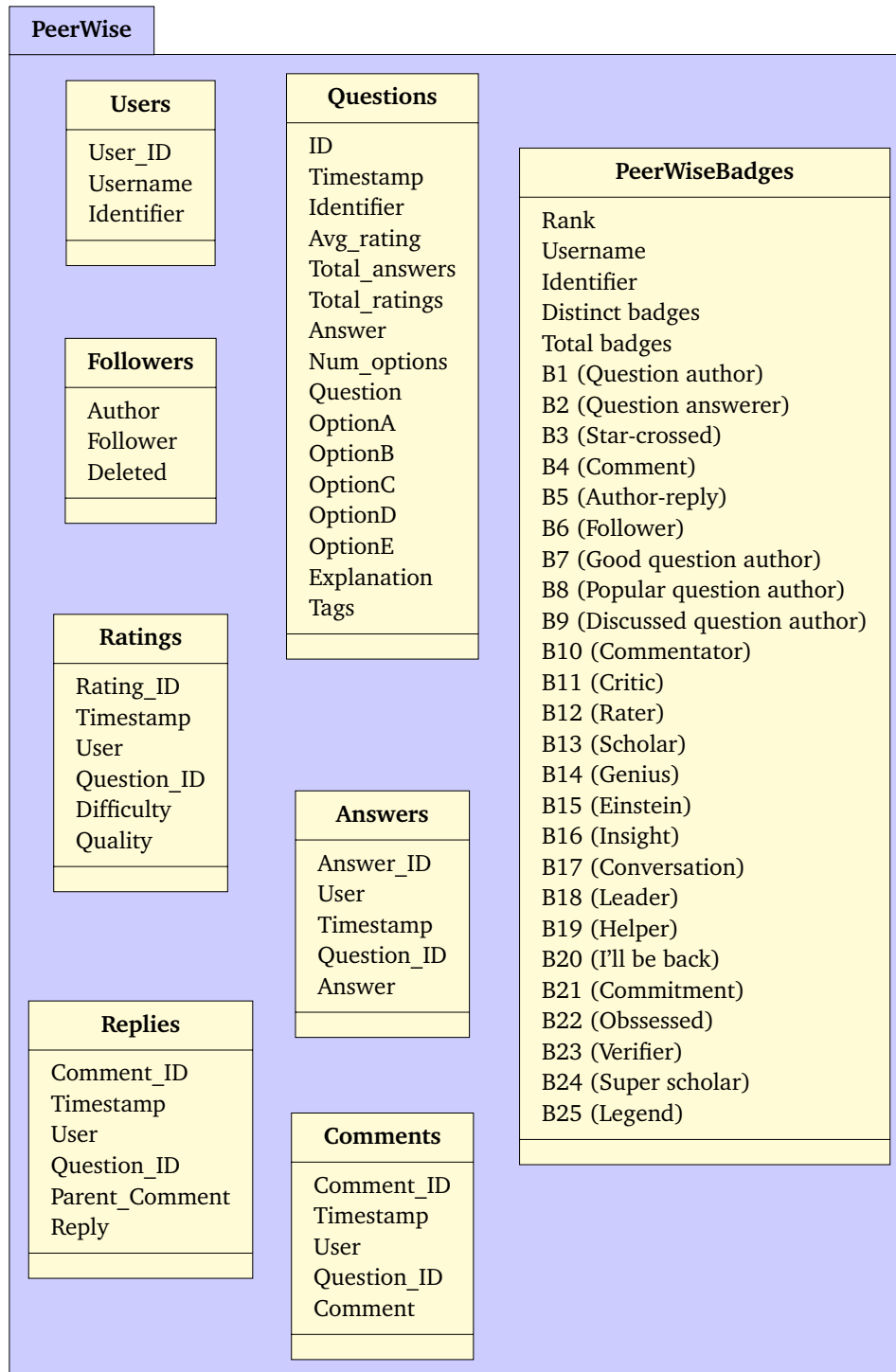
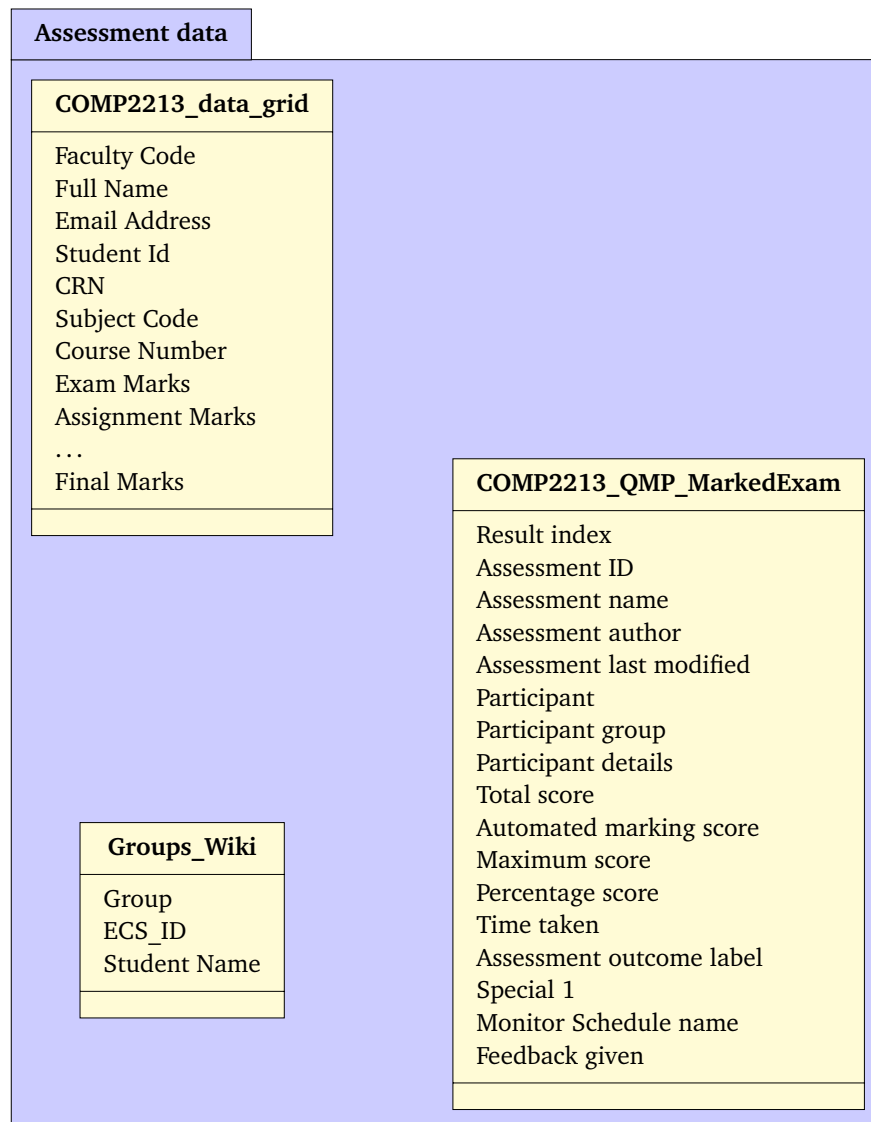




FIGURE 3.6: Associated Files for the PeerWise datasets (II). Some fields are omitted for clarity



for certification eligibility) whereas for learners in the formal context it is the actual marks awarded in their assessment.

### 3.2.3 Clean data

This phase of the feature engineering involved identifying and extracting features from the data, with care about how to deal with duplicates and the data types are consistent and understood as intended<sup>1</sup>.

<sup>1</sup>An example on FutureLearn data is the naming convention for learning steps, being a string of the form "<week number>.<step of the week>". These strings can be mistakenly interpreted as numbers

At this stage there is the opportunity to identify and correct human errors introduced in the data, for example as a result of participants not following exactly the instructions given. This was notably the case for PeerWise data. As shown in the instructions on page F-5 of Appendix F, students had been instructed to use as a “name” their University-of-Southampton username and their student number as the “identifier”. However, some students had instead used nicknames, full email addresses, or their full names as user-name. Yet another student created two different accounts on PeerWise (one with very little activity), so in this phase of the process I manually changed the entries so that the assessment dataset could be successfully merged with the PeerWise data. Subsequently to that step, I then anonymised the clean data and used the unique internal identifier in PeerWise for constructing the feature vector per learner, with those features directly extracted from the data.

Similar challenges were presented with the MOOC data. The files in the datasets were expected to follow strict naming conventions but in reality, some directory names had spaces and unexpected characters (such as parenthesis) that needed to be removed for scripts to work consistently throughout. More critical issues were related to systematically removed data (the unique learner identifier, *learner\_id*, as described in Appendix D), which constituted an unexpected setback of several months of work.

Once data is deemed to be clean and complete, however, problems can still be identified later. For example, the Understanding Language MOOC dataset I received with the learner has an inconsistency I discovered only during the processing of interval activities (explained below) which lead me to believe some entries in the *step-activity.csv* file had been removed either by FutureLearn or by the University of Southampton (perhaps for compliance with data protection). With time at the premium at this stage, I decided to exclude this particular run from subsequent data analysis.

### 3.2.4 Define new features

As mentioned in Section 2.5.1, the design of predictor variables involves lore rather than well-known and validated principles (Baker, 2020). The analysis of the findings obtained through qualitative methods have informed the feature engineering process and the quantitative data analysis.

For each of the courses under consideration I constructed a vector of features (some extracted directly from the log files, other engineered from these) which are representative of the user engagement for each of the peer-supported digital environments within 

---

and those steps with zeroes to the right (e.g. 1.10) would then be mistaken for earlier steps (1.1 in the given example).

which the courses took place. Further, some of the features operationalise concepts that manifest themselves in different ways in each of the environments. To give an example, in the case of MOOCs, certificate eligibility is a commonly used proxy for attainment, whereas in the case of formal instruction, the student attainment is commonly measured through exam marks or final marks.

In order to identify relevant features to engineer from the raw data that each platform collects from the learners, I surveyed the literature seeking understanding on existing work on categorisation of learners based on measures of engagement emerging from the data, which was presented in Sections 2.2 and 2.3. The methodology used for these reviews is detailed in Section ??

Features fall into the following categories:

- Features extracted directly from the datasets
- Derived features (obtained through relatively simple manipulation)
- High-level features (obtained through more complex manipulation)

A non-orthogonal categorisation of features includes:

- Features characterising the learner (e.g. in the formal context: their assessment, and organisational details such as group membership);
- Features capturing temporal information (e.g. when and how often the learner leaves traces of activity, whether they complete tasks sequentially, in an overlapping manner, or not at all);
- Features capturing content production (i.e. engagement with the content);
- Features capturing interaction information (i.e. engagement with each other).

A key driver for feature engineering in this thesis is that they lead to interpretable findings, as informed by a model of learner engagement. This model is presented in full in Chapter 4, but as it represents communicative and non-communicative learning activities occurring in peer-supported digital environment, the previous categorisation could be reformulated as:

- Features characterising the learner's communicative behaviours (e.g. whether they post, make comments or replies, being part of a discussion thread or not);

- Features characterising the learner’s non-communicative behaviours (e.g. how do they engage with the tasks);
- Any other features not directly explainable by this dichotomy, perhaps for presenting elements of both.

### 3.2.5 Deploy

Once a diverse, large feature set was created, I studied those that contribute to meaningful interpretations of patterns of engagement that emerged, by performing attribute selection in the first instance and then a clustering analysis of the data. The effect of these and other engagement features in the learning is then studied using a machine learning technique that involves the grouping of individual data points (clustering), as described in Section 2.5.2.

The design includes the creation of a feature vector associated to each student (made anonymous at source), to characterise their engagement and learning profiles. An important design consideration is that the analysis was to be mirrored with that of a study of learner engagement in the “Understanding language” and “Portus” FutureLearn MOOCs (fully described in section 5.4). This meant that in the engineering of features I was not only guided by choosing those that can be easily derived from the data available, but rather, I became interested on engineering more complex features or at best identifying proxies<sup>2</sup> for those more readily available in the MOOC analysis.

In both cases, a feature vector is constructed to characterise learner engagement with the aim, in the case of MOOCs, of identifying those features which are predictive of retention and completion in the course, rather than of exam performance as in the case of formal settings such as in face-to-face instruction. The research question is whether we can see the same learner behaviour manifesting in different ways in these two peer-supported environments or whether different behaviours altogether are presented.

## 3.3 Summary

This research follows a quantitative methodology, described through the steps in the data science pipeline ranging from data acquisition and curation, to feature extraction and engineering, to the application of a data-driven, unsupervised learning approach on engagement data within MOOCs and PeerWise. In what follows, I will explain in

---

<sup>2</sup>See Definition 1.3 (proxies of learner engagement).

detail the platform agnostic model (in Chapter 4) that was formulated to gain an understanding of this problem, followed by its application and validation in both platforms (in Chapters 5 and 6).



## A platform-agnostic model of learner interactions

*“I learned without any pressure of punishment to urge me on, for my heart urged me to give birth to its conceptions, which I could only do by learning words not of those who taught, but of those who talked with me; in whose ears also I gave birth to the thoughts, whatever I conceived.”*

St Augustine (b.354–d.430), “[THE CONFESSIONS OF SAINT AUGUSTINE](#)”, AD 401.

Discussions and conversations are effective catalysts for learning, as illustrated in this epigraph<sup>1</sup>. St Augustine, one of the most prolific authors from ancient times, devoted much of his writing to considerations around knowledge, teaching and learning. He saw teaching as a mere preparatory mechanism for understanding, and considered conversations as the true key to unlock the power of learning. Only relatively recently (just over a century ago), the contrasting concept of “learning by doing” became more prevalent through the work by John Dewey and others, discussed in Section [2.1.4](#).

I take these two ideas to motivate the development of a platform-agnostic model of learning engagement within peer-supported environments that is defined in this chapter. I do so by identifying *conversing* and *doing* activities in these environments as complementary in facilitating learning. The way they are complementary is central to the formulation of the new platform-agnostic model of learner interactions introduced in the previous chapter.

---

<sup>1</sup>Often paraphrased as “I learned most, not from those who taught me, but from those who talked with me”.

More specifically, in this chapter I describe in detail the interactions modelled, which include those amongst learners (conversing) and those of learners with the learning material (doing). The distinction between communicative activities (those involving conversations with peers) and non-communicative activities (those where the learner is engaging with the material) was already made as part of the high-level overview of the methodology in this thesis, in Chapter 3. The essential interaction processes within peer-supported digital environments are shown in Figure 4.1.

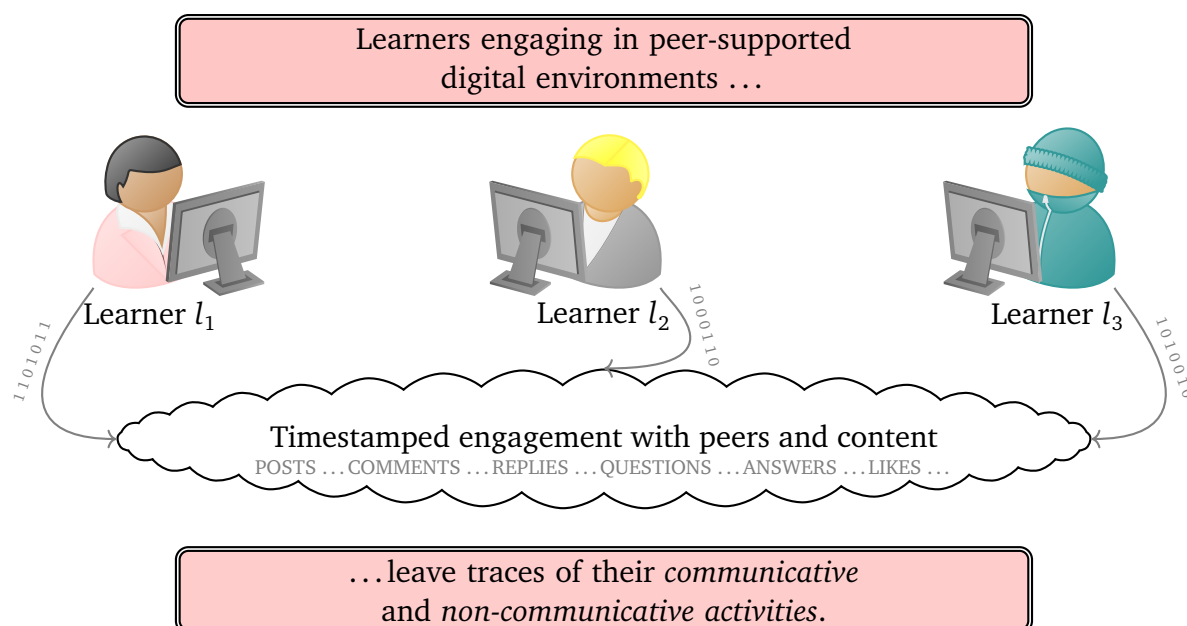


FIGURE 4.1: Basis of the platform-agnostic model of learner interactions (adapted from Figure 3.1).

This chapter concerns **RQ1**: *In the context of peer-supported digital environments, is it possible to define a platform-agnostic model of learner interactions (with the content matter and with each another)?* To answer this research question, I formulate a theoretical model that, independently of the platform implementation, conceptualises learner activities, given their observable digital “traces”.

To illustrate the difference between activities and their traces, consider the following situation: suppose a dancer performing behind a translucent screen, so that observers looking at the screen cannot see the dancer directly, but only as a silhouette on the screen. Though the observers are aware that the dance takes place in a three-dimensional space, some of it is completely occluded (e.g. when dancing behind opaque objects) and all that is observable is a two-dimensional silhouette on the screen, as illustrated in Figure 4.2.

Silhouettes reflect, though they are not able to fully convey, the richness of the real



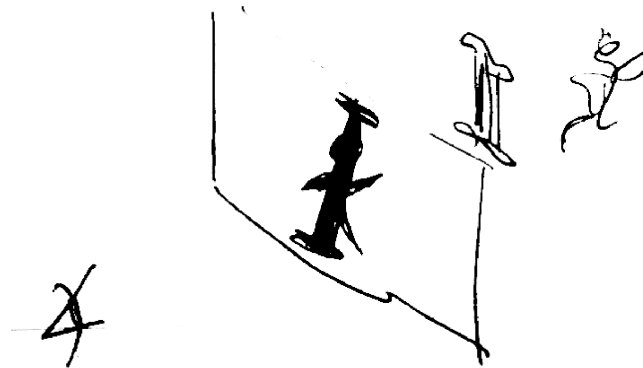


FIGURE 4.2: The silhouette of a partly-occluded dancer performing behind a translucent screen represents the visible traces of activity. The performance represents the learning and the dancer, the learner.

activity of which they are mere traces. In digital learning environments, the equivalent to these ‘silhouettes’, are digital traces, also referred to as “microanalytics” in the context of some educational software, such as Talis (Dix, 2016). In what follows, these will be referred to as *e-tivities*, as a portmanteau of *electronically-captured activities*. I borrow this term from Salmon (2002), who used it to describe high-level design plans for learning activities, as discussed in Section 2.1.3 of Chapter 2.

However, in this context it is used to mean the recorded evidence of a learning activity of *fine granularity* that has taken place within the digital environment, be it of the communicative kind or not. This finer granularity for e-learning activities allows for the capture of those considered to be of an atomic nature (i.e., those activities of such short duration that are either completed by the learner at a given point in time or not attempted at all and thus not recorded), but also those which span periods of time, with a determined beginning and a possibly also an end. I define e-tivities precisely in Section 4.1, and elaborate the model through the subsequent Sections, particularly: types of e-tivities (in Section 4.2), before detailing further the types of communicative e-tivities (Section 4.3) and non-communicative ones (in Section 4.4). Having completed the formalisation of the model, I outline its limitations in Section 4.5. Finally, I offer a summary of the full model and suggest a validation strategy in Section 4.6.

## 4.1 e-tivities

An *electronically-captured activity*, or *e-tivity* within a peer-supported digital environment, is here defined as a triple representing the relationship between a learning activity, the learner who performed it, and the specific time in which (or period during which) it

took place. More precisely, the triples  $\langle a_i, l_j, t_k \rangle$  and  $\langle a_i, l_j, d_k \rangle$  would indicate that the activity  $a_i$  (the  $i$ -th distinct fine-granularity activity recorded within the environment) was performed by learner  $l_j$  (the  $j$ -th enrolled learner) at a specific moment in time  $t_k$  for the former, or during a period  $d_k$ , for the latter. In other words, the answers to three key questions: ‘what?’, ‘who?’, ‘when?’ about the e-tivity, as follows:

**what** Any of the prescribed learning activities supported by the platform, more formally  $a \in A$ , the set of *activities*. These activities involve either communication between learners or engagement with the learning material, such as production or consumption of content. This distinction may be dynamic and highly context-dependent, as I will discuss in Section 4.2.

**who** Typically *learners* in a course, though this model could well include other actors in the learning activity, such as educators and moderators who may act as if they were ‘peers’ engaging with others and the learning materials, as discussed in Section 2.1.4 in Chapter 2. More formally,  $L$  is the set of learners on a course, such that  $|L|$  is the cohort size.

**when** Either a specific point in time, typically characterised by a timestamp  $t$  in a computer implementation, or a period  $d$  which has a starting time  $t_s$ , and possibly an end time  $t_e$ , with  $T$  being the set of timestamps with a total order  $<$ . The total order  $<$  entails that, for any given two timestamps, one either precedes or succeeds the other. Further, a timestamp with a given sub-index will always precede timestamps of a higher sub-index. Timestamps and intervals (or periods) will be further discussed in Section 4.4.1.

Let us consider the following definitions that are the basis of this model:

**Definition 4.1** (Activities). The set  $A$  is the set of all fine-granularity *activities*  $a_i$  that can be captured in a given peer-supported digital environment, such that  $a_i$  is either a monuple  $w_i \in A_N$  or a pair  $\langle c_i, m_j \rangle \in A_C$ . Therefore:

$$A = A_C \cup A_N \quad (4.1)$$

$$A_C \cap A_N = \emptyset \quad (4.2)$$

**Definition 4.2** (Non-communicative activities). The set  $A_N \subseteq A$  is the set of non communicative activities  $w_i$  such that  $w_i$  is assumed to be some work undertaken within a given peer-supported digital environment over a period of time (i.e. not primarily intended for communication).

$$A_N = \{w_1, w_2, \dots, w_n\} \quad (4.3)$$

**Definition 4.3** (Communicative activities). The set  $A_C \subseteq A$  is the set of communicative activities  $\langle c, m \rangle$  where  $c \in M$  is a communication in response to a message  $m \in M_0$ .  $M$  is the set of messages, and  $M_0$  is the set of messages that includes the distinguished member  $m_{\perp}$ , the *null* message (not to be confused with an empty message).

$$A_C = \{\langle c, m \rangle \mid c \in M \wedge m \in M_0\} \quad (4.4)$$

Note that  $\forall i, j : \langle c_i, m_j \rangle$ , these are instantaneous, intentional communications.

**Definition 4.4** (e-tivities). The set  $E$  is the set of electronically-captured activities *e-tivities* comprising all of the activities  $a_i \in A$  undertaken by each learner  $l_j \in L$  at a time  $t_k \in T$  or over a period of time  $d_k \in D$ , such that  $\forall t_i, t_j \in T : \text{if } i < j \text{ then } t_i < t_j$ .

$$E = \{\langle a, l, t \rangle \mid a \in A, l \in L, t \in T \cup D\} \quad (4.5)$$

## 4.2 Types of e-tivities

As established in the motivation earlier in this chapter, when studying learner interactions I am interested the distinction between conversing and doing, and how learners engage with each other and with their learning content through communicative and non-communicative activities. The key aspects informing my modelling are listed in Table 4.1 with my view on how these activities are complementary.

TABLE 4.1: Complementary interpretations of views on learning activities

Communicative activities		Non-communicative activities
Conversations with others	Centered around	Individuals' work
Conversing	Learning paradigm	Doing
Typically instantaneous	Time per activity	Typically over a period
Primarily interaction	Focus on	Primarily knowledge and skills
Others	Interaction with	Content
Outwards into the external world	Knowledge "flow"	Inwards into oneself
Socrates, St Augustine, Pask, Laurillard, Baker	Known advocates	Dewey, Kolb

In practice, these will be any of the prescribed learning activities that may be afforded via any of the platforms for peer-supported digital learning environments, as per the following non-exhaustive list:

- asking questions
- posting comments
- giving replies
- reading comments
- reading content
- creating videos
- watching videos
- giving “likes”
- give a quality/difficulty rating
- writing essays
- writing a peer-review on someone else’s essay
- creating multiple-choice questions
- answering multiple-choice questions
- filling the blanks in a cloze exercise
- answering surveys
- following “posters” (those who post content)

Sometimes it is challenging to make the distinction between what makes an activity communicative or not. This is because even when the purpose of the learning activity may not be primarily communicative, when the activity is visible to others it may spark conversations, as it is often the case with learner-generated content. Learners who engage with their material in an active way, say, **producing** content in a peer-supported digital environment, implicitly offer that product to the scrutiny of others, who in turn may engage with it by **conversing** about it or by **consuming** it, as shown in Figure 4.3.

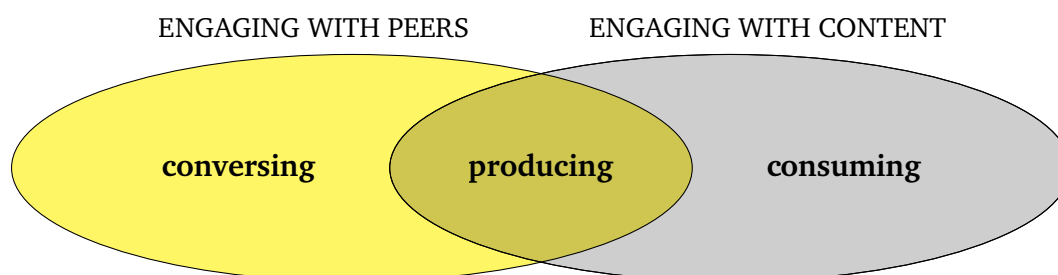


FIGURE 4.3: Engagement within a peer-supported digital environment. Learners engage with others through *communicative* activities. Passive engagement with the material (consuming it) is *non-communicative*, as produced content often sparks conversations.

A good example is the creation of multiple-choice questions (MCQs), which becomes a resource for others to engage with in two different ways: with the associated learning content (since answering the MCQ would be a way to consume it), and with the learner who created it (making a comment, hence continuing a conversation about it).

Still, the Augustinian/Deweyan inspired views for learner engagement illustrated above can be developed further. Conversing is by nature an *active* function, requiring at least two parties to make contributions, and therefore engaging with peers can be

seen in theory as a purely active type of engagement. Indeed, as per the epigraph of this chapter, St Augustine is said to have “*learned most, not from those who taught*” him (suggesting a passive engagement), “*but from those who talked **with***” him (the emphasis is mine, to highlight the active participation of both parties).

Nowadays however, many peer-supported digital environments support various affordances that allow for a *passive* kind of engagement with peers: following others, reading their comments or liking their contributions, for example. These are all forms of engagement with others that tend to be much more passive than, for example, asking questions, writing comments or giving replies. The main difference between these two ways of engagement with *peers* (whether it is active or passive) can perhaps be seen more evidently in activities fostering engagement with the learning *content*. We could classify that engagement as either active or passive depending on whether the content is being produced or consumed by the learner. Perhaps more accurately however, given that peer-supported digital environments often offer ways to engage with others somewhat passively, learner engagement could be placed somewhere along the active/passive spectrum. This would depend on how active the production act is (or how passive is the consumption of the material). For example, “making an MCQ” is more active than “answering an MCQ” even though in both learners are producing something, it can be seen that answering the MCQ is somewhat an act of consumption of the MCQ that was previously produced by someone else. Similarly, “reading” a comment is a more passive act than “liking” it, but both could be regarded as observing acts.

Therefore in addition to the three kinds of learner engagement illustrated in Figure 4.3 (conversing, producing and consuming), as discussed, I include “**observing**”, to reflect that the learner interactions with each other could be ‘consumed’ passively, just as well as the learning content can be.

Some of the examples listed in this discussion are shown across the two-dimensional space in Figure 4.4, with the dimensions being whether the engagement is active or passive and whether it is with peers or the learning content, with the caveat that the quadrants **producing** and **conversing** are rather fluid as discussed, depending on the context as products of activities may be the start of conversations.

Given the above discussion about communicative vs non-communicative activities, let us remember that e-tivities are triples comprising “what” (the activity), “who” (the learner), and “when” (the time), as per Definition 4.4. This means that e-tivities can be either communicative or non-communicative too, according to what type of activity is being captured as the first element in the triple. Therefore, *communicative e-tivities* are those used for communication with peers, by a learner, at a given time. Examples would be:  $\langle \text{a comment, Ana, } t_0 \rangle$  and  $\langle \text{a reply, Bob, } t_1 \rangle$ . In contrast, *non-communicative e-tivities*

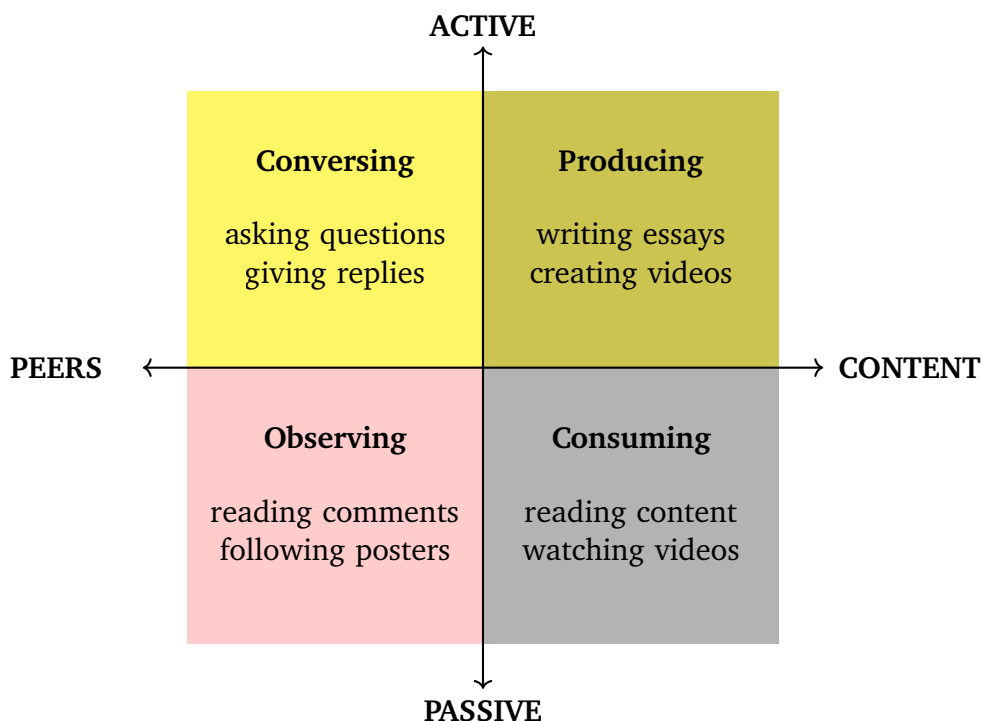


FIGURE 4.4: Examples of typical learning activities in various digital environments and where they lie in the “conversing vs doing” space, with the dimensions being: who/what are they engaging with, and the whether their engagement is active or passive.

are those in which the learner is engaging with the learning material as an individual, either consuming it (such as watching a video, or answering an MCQ) or producing it. Sometimes content produced by a learner, though originally as the product of a non-communicative activity, can become part of a conversation (if it attracts comments from others, for example) and therefore the associated e-tivity becomes communicative. The implication for this is that membership to the sets

More formally, a partition  $P_E$  of  $E$  into these two sets satisfies:

**Definition 4.5** (Partition of  $E$ ). The set of e-tivities  $E$  is partitioned as two sets:  $E_C$  (comprising all communicative e-tivities) and  $E_N$  (non-communicative e-tivities), such that:

$$(E = E_C \cup E_N) \quad \wedge \quad (E_C \cap E_N = \emptyset) \quad (4.6)$$

Next, in Section 4.3 and Section 4.4, I develop the model of learner engagement within peer-supported digital environments through each of these two types of e-tivities.

### 4.3 Communicative e-tivities

As mentioned, the difference between communicative and non-communicative activities can be merely contextual, i.e. its type will depend on whether related e-tivities occur later or not. Understanding the relevance of capturing the context where learner contributions take place is fundamental to the model of interactions I am proposing. This is inspired by the turn-taking nature of dialogues observed by [Chua et al. \(2017\)](#), discussed in Chapter 2. When the dialogic context is taken into account, communicative e-tivities can be broadly classified as one of the following five types:

**SP** *starting posts* (communications by a learner at a given time that are responded to by others at a later time),

**LP** *lone posts* (learner contributions that are not responded to by others, even though the learner may have added further information in a later post as a “reply to self”),

**FR** *first replies* (responses to a *starting post* that had been communicated by another learner at an earlier time),

**IR** *initiators’ replies* (responses to others’ replies to one’s own starting post), and,

**AR** *additional replies* (responses to others’ replies under a starting post that has already been replied to).

More formally:

**Definition 4.6** ( $P_{Ec}$ ). The partition  $P_{Ec}$  of  $E_C$  is the set  $P_{Ec} = \{E_{SP}, E_{LP}, E_{FR}, E_{IR}, E_{AR}\}$  satisfying:

$$E_C = E_{SP} \cup E_{LP} \cup E_{FR} \cup E_{IR} \cup E_{AR} \quad (4.7)$$

$$\begin{array}{llll} E_{SP} \cap E_{LP} = \emptyset & & & \\ E_{SP} \cap E_{FR} = \emptyset & E_{LP} \cap E_{FR} = \emptyset & E_{FR} \cap E_{IR} = \emptyset & \\ E_{SP} \cap E_{IR} = \emptyset & E_{LP} \cap E_{IR} = \emptyset & E_{FR} \cap E_{AR} = \emptyset & E_{IR} \cap E_{LP} = \emptyset \\ E_{SP} \cap E_{AR} = \emptyset & E_{LP} \cap E_{AR} = \emptyset & & \end{array}$$

where  $E_{SP}$  is the set of all starting posts,  $E_{LP}$  is the set of all lone posts,  $E_{FR}$  is the set of all first replies,  $E_{IR}$  is the set of all initiators’ replies, and  $E_{AR}$  is the set of all additional replies. In other words, each of the communicative e-tivities  $e = \langle a, l, t \rangle \in E_C$  belong to one and exactly one of these five sets  $E_{SP}, E_{LP}, E_{FR}, E_{IR}$  or  $E_{AR}$ .

The belonging to one of these five sets is determined by a contextual function  $F$ , defined as follows:

**Definition 4.7** (Response-to). The *response-to* function  $F : E_C \rightarrow E_C$  maps a communicative e-tivity  $e$  to another one  $f$ , such that  $f$  is a direct response to  $e$ , and no other e-tivity. In other words,  $F(f) = e$ , or, more precisely:

$$F \subseteq \{ \langle \langle a, b \rangle, l, t \rangle, \langle \langle b, c \rangle, l', t' \rangle \mid a, b, c \in M \wedge l, l' \in L \wedge t, t' \in T \} \quad (4.8)$$

The function  $F$  is irreflexive, antisymmetric and non-transitive<sup>2</sup>.

$F^+$ , the transitive closure of  $F$ , relates indirectly a communicative e-tivity to all of its ancestors, so for example  $F^2(e_2) = F(F(e_2)) = e_0$  means that there exists a communicative e-tivity  $e_1$  such that  $F(e_2) = e_1$  and  $F(e_1) = e_0$ .

Similarly,  $F^n(e_n) = F(F(\dots F(e_n)\dots)) = e_0$  means that there exist communicative e-tivities  $e_{n-1}, e_{n-2}, \dots, e_1$  such that  $F(e_n) = e_{n-1} \wedge F(e_{n-1}) = e_{n-2} \wedge \dots \wedge F(e_1) = e_0$ .

**Definition 4.8** (Starting post). The communicative e-tivity  $e$  is a *starting post* if there exists a communicative e-tivity  $f$  such that  $f$  is a direct response to  $e$  not created by the same learner.

$$E_{SP} = \{ e = \langle a, l, t \rangle \mid \exists f = \langle a', l', t' \rangle \in E_C \wedge F(f) = e \wedge l \neq l' \} \quad (4.9)$$

**Definition 4.9** (First reply). The communicative e-tivity  $f$  is a *first reply* if there exists a starting post  $e$  such that  $f$  is a direct response to  $e$  not created by the same learner.

$$E_{FR} = \{ f = \langle a, l, t \rangle \mid \exists e = \langle a', l', t' \rangle \in E_{SP} \wedge F(f) = e \wedge l \neq l' \} \quad (4.10)$$

**Definition 4.10** (Initiator's reply). The communicative e-tivity  $f$  is an *initiator's reply* if there exist a communicative e-tivity that is a starting post  $e \in E_{SP}$  and a positive integer  $k > 1$ , such that  $F^k(f) = e$  (i.e.  $f$  is an indirect response to  $e$ ), and with  $e$  and  $f$  having both been produced by the same learner. More precisely:

$$E_{IR} = \{ \langle a, l, t \rangle \in E_C \mid \exists \langle a', l', t' \rangle \in E_{SP} \wedge \exists k \in \mathbb{N}, k > 1 \wedge F^k(\langle a, l, t \rangle) = \langle a', l', t' \rangle \} \quad (4.11)$$

<sup>2</sup>The properties of  $F$  as per the Definition 4.7 mean that an e-tivity  $e$  cannot be its own reply (i.e.  $\langle e, e \rangle \notin F$ ), and that if  $f$  is a reply of  $e$ , then  $e$  cannot be a reply of  $f$ . Also, if an e-tivity  $e$  is a direct reply of  $f$ , and  $f$  a direct reply of  $g$ , then it does not follow that  $e$  is a direct reply of  $g$ .



**Definition 4.11** (Additional reply). The communicative e-tivity  $f$  is an *additional reply* if there exist a communicative e-tivity that is a starting post  $e \in E_{SP}$  and a positive integer  $k > 1$ , such that  $F^k(f) = e$  (i.e.  $f$  is an indirect response to  $e$ ), and with  $e$  and  $f$  having been produced by different learners.

$$E_{AR} = \{ \langle a, l, t \rangle \in E_C \mid \exists \langle a', l', t' \rangle \in E_{SP} \wedge (\exists k \in \mathbb{N}, k > 1) [F^k(\langle a, l, t \rangle) = \langle a', l', t' \rangle] \wedge l \neq l' \} \quad (4.12)$$

**Definition 4.12** (Discussion thread). The *discussion thread* function  $DT : E_C \rightarrow E_C^*$  maps a communicative e-tivity  $e_k$  with the sequence of communicative e-tivities  $e_i$ , such that each  $e_i$  is a response to another e-tivity in the sequence. This sequence includes  $e_k$  and the posts to which  $e_k$  is a direct or indirect response to (upto and including the post  $e_0$  from which the thread of responses is originated). More formally:

$$DT(e_k) = \left\{ \langle e_0, e_1, \dots, e_k \rangle \mid F(e_0) = m_{\perp} \wedge \left[ \bigwedge_{i=1}^k F^i(e_i) = e_0 \right] \right\} \quad (4.13)$$

where  $m_{\perp}$  is the distinguished member of  $M^*$ , the *null* message.

**Definition 4.13** (Lone post). The communicative e-tivity  $e \langle a, l, t \rangle$  is a *lone post* if it is not part of any discussion threads involving any learners other than  $l$ . More formally, if there does not exist an e-tivity  $f = \langle a', l', t' \rangle$  in the discussion thread of  $e$ , such that  $l \neq l'$ , and, it does not appear in the discussion thread of any e-tivities involving other learners.

$$E_{LP} = \{ \langle a, l, t \rangle \mid \nexists \langle a', l', t' \rangle [DT(\langle a, l, t \rangle) = \langle a', l', t' \rangle \wedge \langle a, l, t \rangle \notin DT(\langle a', l', t' \rangle)] \wedge l \neq l' \} \quad (4.14)$$

In other words, all e-tivities (if any) in the discussion thread are produced by the same learner: there is no ‘conversation’ with other learners.

To illustrate how these definitions are the basis of a platform-agnostic model of learner engagement within peer-supported digital environments that can be applied in practice, the following section offers a simple hypothetical scenario.

### 4.3.1 A practical scenario

Consider a hypothetical scenario where various contributions and conversations amongst learners take place in a peer-supported digital environment. Here, three learners, Ana,

Bob and Cam, post some comments, which, as I will show, correspond to different types of e-tivities, given the context in which they occur.

Figure 4.5 shows a chat-like history of comments produced by each of these three learners, as follows: comments by Ana are shown on the left, in **pink** speech bubbles, those by Bob in the middle, in **grey**, and those by Cam on the right, in **teal**.

The temporal relationship amongst the comments posted is shown vertically, with earlier comments appearing at the top and later comments added underneath, until the last comment at the bottom of the figure. Though no threading or nesting is shown (and indeed this would be an implementation detail, platform-dependent, which may or may not be supported), it is evident that some comments stand alone, unanswered, whilst others may elicit responses, even if not immediately and many other events may have occurred in between.

Scenarios such as the one illustrated could be represented in a different way in order to capture the contextual relationships amongst posts and their replies. One way would be to use a forest of trees rooted in each of the learners at play with their initiating (or lone) posts shown in the first level, the first replies received on their starting posts in the second level, and their additional replies appearing deeper in their respective trees. The rationale behind applying such a structure is that each learner could be characterised by the kind of tree it produces.

For example, a very deep tree would indicate that the learner tends to initiate longer interactions than, for example, another learner whose associated tree is very shallow. Also, the breadth of the tree could be an indicator of how many distinct conversations the learner initiates. Further, being able to organise learners' communicative e-tivities in such manner, facilitates the visualisation the kinds of e-tivities created. In particular, whether they stand alone as starting posts (or are "zero-order replies"), direct responses to starting posts ("first-order replies"), or indirect ones ("second-order" replies)<sup>3</sup>. In such a structure, the more direct responses to starting posts would be placed closer to the root of each tree, regardless of the time when were they posted. The insights provided by these kinds of observations about learners and how they engage with each other can be useful to understand learner behaviours. Therefore, this model can inform feature engineering processes for data analysis of learner data on specific platforms, as shown in Sections 6.4 and 5.4 where I describe the features extracted in PeerWise data and MOOCs data, respectively.

In the particular case of the scenario in Figure 4.5, applying such a process results

---

<sup>3</sup>The formal definitions of zero-, first- and second-order replies are Definitions 4.14, 4.15 and 4.16 respectively).



FIGURE 4.5: Hypothetical scenario with a set of conversations amongst learners.

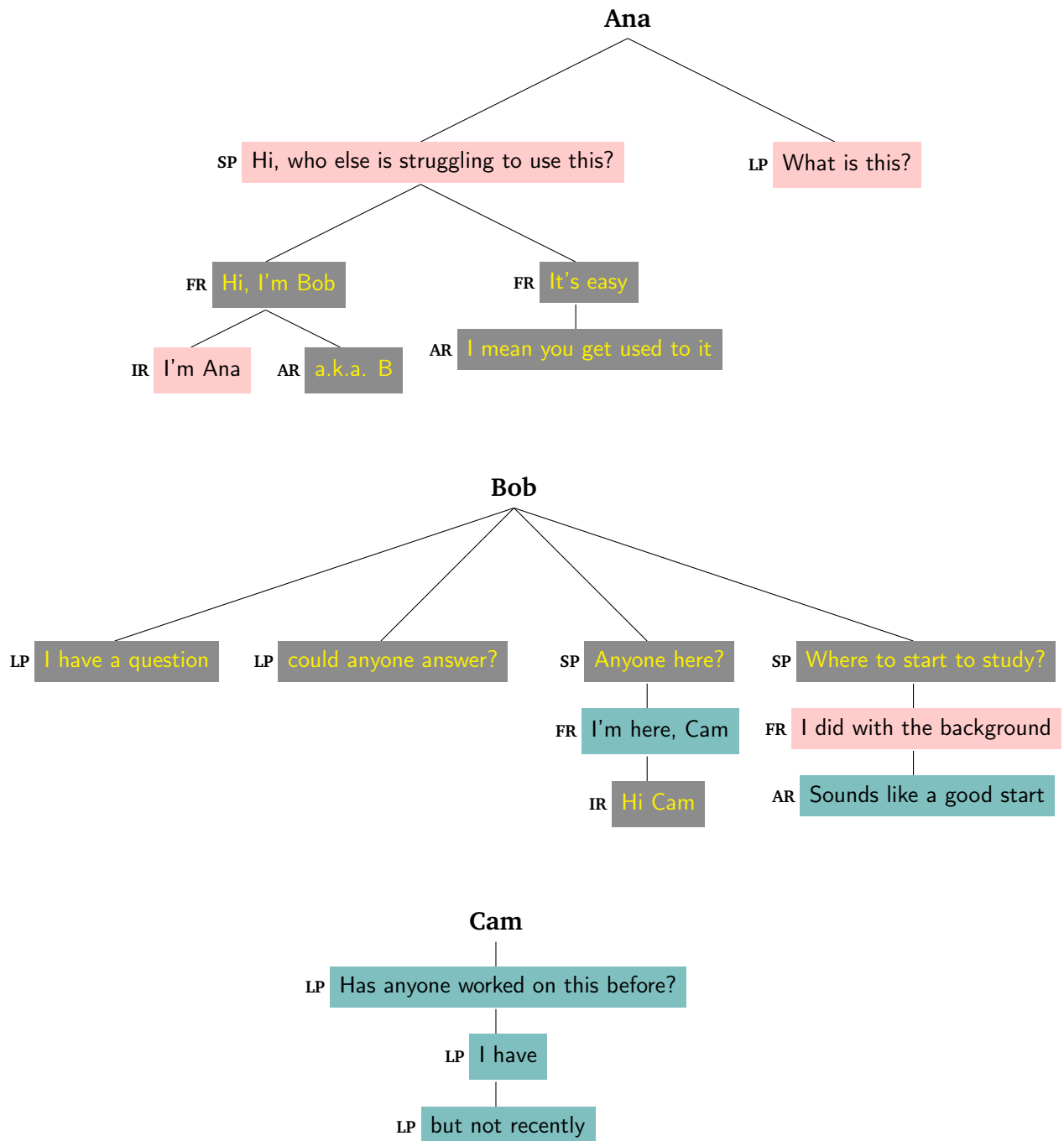


FIGURE 4.6: Alternative representation of the illustrated scenario, showing the contextual relationships between learners and their posts, comments and replies. To the left of each post, a code is given to indicate their types of communicative e-tivity: starting post (SP), lone post (LP), first reply (FR), additional reply (AR) and initiator's reply (IR).

in the tree shown in Figure 4.6. For example, Ana’s start of her contributions “Hi, who else is struggling to use this?”, elicits a greeting back from Bob immediately after, “Hi, I’m Bob”, making her comment a **starting post** in the conversation. Then, her reply, “I’m Ana” becomes, due to this context, an **initiator’s reply**. Bob’s **additional reply**, “a.k.a. B”, continues the conversation. By contrast, her second post, “What is this?”, goes unanswered during the whole of the recorded exchanges, therefore remaining to be a **lone post**. Indeed, several **lone posts** were created by all, not just Ana, but also Bob (“I have a question”, “could anyone answer?”) and Cam (“Has anyone worked on this before?”, “I have” and “but not recently”). Notably, whilst Cam is seemingly answering his own question, this is not a conversation but a succession of lone posts, given that these do not involve other learners. **First replies** are perhaps more intuitive to identify, namely “Hi, I’m Bob”, in answer to Ana’s post, and “I’m here, Cam” in answer to Bob’s.

In contrast, Bob’s earlier comment, “It’s easy”, is actually a first reply to the second part of Ana’s first post “Hi, who else is struggling to use this?”, to which only much later he gives a **additional reply** by saying “I mean you get used to it” (the last comment listed). In the interim, other conversations had started (or not).

### 4.3.2 Limitations of the chat representation

Though the multi-column, multi-colour chat-like representation shown in Figure 4.5 and the multi-coloured trees in Figure 4.6 both allow to visualise the actors at play and their contributions (over time and in context, respectively), a general model, capable of capturing a much larger number of learners interacting, cannot rely on the use of colour, which I have used so far for illustrative purposes. Ultimately, given that the model of learner engagement within peer-supported environments is based on e-tivities, which are triples of the form ⟨“what”, “who”, “when”⟩ (as per the definition given in Section 4.1), it is necessary to make a mapping that incorporates explicitly each of the three elements in these triples.

Let us first consider “who” and “when”. The “who” are the *learners*, Ana, Bob and Cam in our hypothetical scenario, who now become  $l_1$ ,  $l_2$  and  $l_3$  in the model (the first three out of a possible  $|L|$  number of learners in the cohort). The “when” would be the timestamps associated to the vertical timeline intuitively suggested visually in the chat-like representation of Figure 4.5, satisfying that  $\forall i < j : t_i < t_j$  (i.e. that the indices of all timestamps follow the same ordering than the actual timestamp values). Applying such a mapping to the hypothetical scenario above described (the chat-like representation), with the identified types of communicative activities (starting post, lone post, first reply, initiator’s reply and additional reply, as per the mapping shown in the colourful tree in

Figure 4.6), would result in the sequence of e-tivities shown in Figure 4.7.

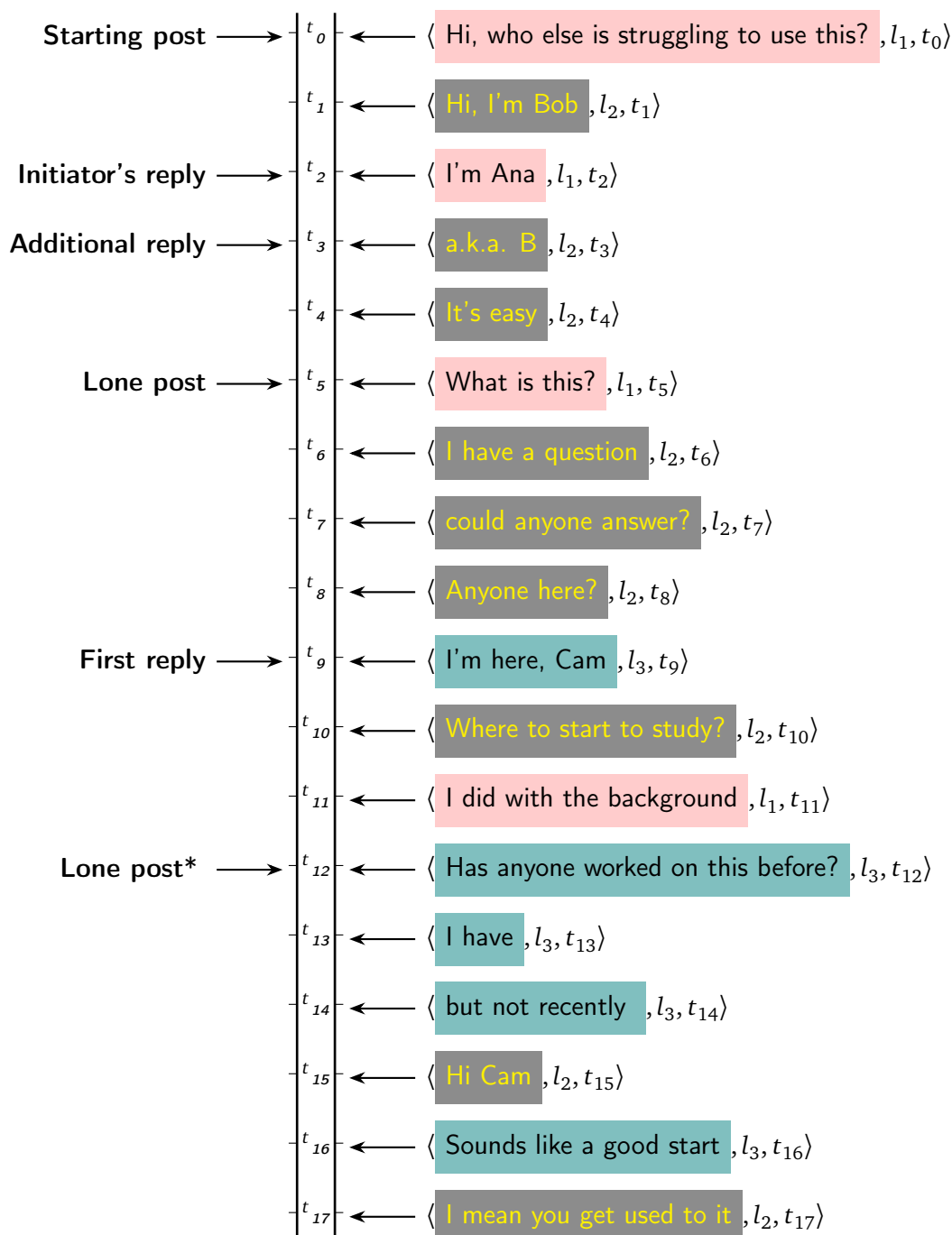


FIGURE 4.7: Timeline of posts, comments and replies in the scenario introduced in Figure 4.5, showing various types of interactions, and the times when they took place.

\*Note that post in  $t_{12}$  is still a **lone post**, since even though it has sparked a comment and a reply, these are from the initiator,  $l_3$ , Cam.

Now the information about the learners and the timestamps of the posts are incorporated (“who” and “when”), I turn my attention to “what”. The general model of learner engagement requires to make abstraction of the utterances exchanged to focus on the

contextual nature of the role they played in the conversations when they occurred, e.g. whether they were in response to others. This with the ultimate goal of replacing the texts with their corresponding abstractions as shown in the following section.

### 4.3.3 Communicative e-tivities as n-order replies

To model learner interactions within a peer-supported digital environment, utterances can be replaced with an identifier that makes an abstraction of their role in the dialogue. Suitable identifiers, such as *posts*, *comments* and *replies*, could be assigned to communicative e-tivities based on whether they are zero-, first- or second-order replies.

**Definition 4.14** (Zero-order reply). A zero-order reply, or *post* e-tivity, comprises any communicative activity (usually “posts”) created not in response to a previously recorded one. The ‘zero’ means to emphasise that it is *not* a reply, i.e.  $P = \{p \in E_C \wedge F(p) \notin E_C\}$ .

It is made by a learner  $l$  at a moment in time of timestamp  $t$ . Since all learners can make many posts, each of their posts will be uniquely identified as  $p_j$ , representing the  $j^{th}$  post recorded in the peer-supported digital environment, which could have been made by any learner.

Therefore, let  $P$  be the set of *posts*:

$$P = \left\{ \langle \langle p_j, m_{\perp} \rangle, l, t \rangle \mid p_j \in M, m_{\perp} \in M^*, l \in L, t \in T \right\} \quad (4.15)$$

where  $m_{\perp}$  is the distinguished member of  $M^*$ , the *null* message.

Only *starting posts* and *lone posts*<sup>4</sup> can be found under this category, and correspond to all nodes at depth=1 in the tree of e-tivities, as in the example shown in Figure 4.6. Starting posts would have sparked comments by other learners at depth=2, whereas lone post would either be leaves at this level (i.e. no comments associated) or have comments made by the same learner who created the original post.

**Definition 4.15** (First-order reply). A first-order reply, or *comment* e-tivity, comprises all communicative activities (usually “comments”) that are in response to a zero-order reply, i.e.  $C = \{c \in E_C \wedge F(c) \in P\}$ .

In other words, comments made by any learner  $l_i$  at any moment of timestamp  $t_k$  on every zero-order reply, such as a post  $p_j \in P$ . Since all learners can make many

<sup>4</sup>However, not all lone posts are zero-order replies, as for example the lone post in timestamp  $t_{13}$  in Figure 4.7. As indicated in the caption, the post in  $t_{12}$  is a lone post which has a reply. However, this is a “reply to self”, and as such, it is also considered in the model as a lone post, because it is not part of a conversation, irrespective of its depth in the tree.)

comments on a given post  $p_j$ , each of their comments will be uniquely identified as  $c_n$ , representing the  $n^{\text{th}}$  comment recorded in the peer-supported digital environment, which could have been made by any learner. In order to capture what is the post the comment is a reply to, the activity  $a$  in the e-tivity tuple  $\langle a, l, t \rangle$  is represented as the tuple  $\langle c_n, p_j \rangle$ .

Therefore,  $C = \langle \langle c_n, p_j \rangle, l_i, t_k \rangle$  is the set of comments such that  $F(c_n) = p_j$  and  $p_j \in P$ .

Under this category we can find only *first replies* and *lone posts* (in particular, those lone posts that are ‘replies to self’), and correspond to all nodes at depth=2 in the forest of e-tivities. These posts are direct replies to starting posts (and therefore will be first replies, provided the posters are not commenting upon posts made by themselves, in which case these are *lone posts*, as shown).

**Definition 4.16** (Second-order reply). A second-order reply, or *reply e-tivity*, comprises a reply  $r$  made by learner  $l_i$  at a moment in time of timestamp  $t_k$  to an activity  $a$ , which could be: either a first-order reply  $c_n$  (related to post  $p_j$ , as above), or another second-order reply  $r_{n'}$  (related to a comment  $c_n$ ).

Since all learners can make more than one reply to a given comment  $c_n$  or a previous reply  $r_u$ , each of these replies will be uniquely identified as  $r_v$ , representing the  $v^{\text{th}}$  reply recorded in the peer-supported digital environment, made by any learner. Note that  $u < v$  is maintained as an invariant, in other words, whilst replies can be given to other replies, these must already exist by definition, thus, their sub-index  $u$  will always be smaller than the sub-index of the new reply, which represents that the reply  $r_u$  was added to the set  $R$  prior to the new reply  $r_v$ , and this is true for all replies  $r_u$  and  $r_v \in R$ .

Let  $R$  be the set of replies:

$$\langle \langle r_n, a \rangle, l_i, t_k \rangle \ni \left( (a = c_k \wedge c_k \in C) \vee (a = r_m \wedge r_m \in R) \right) \wedge \\ F(r_n) = a \wedge m, n = 1..|R| \wedge m < n$$

Under this category we find *initiators' replies* and *additional replies*, and correspond to all nodes of depth  $> 2$  in the tree of e-tivities, which suggests it is part of a conversation including indirect replies to a starting post (either by the initiator or by others, provided that in the conversation there has been more than one learner involved, otherwise, like in previous cases, these will be *lone posts*).

The formal definitions of zero-, first- and second-order replies given above provide an improved way of representing the hypothetical exchange in this scenario, that is



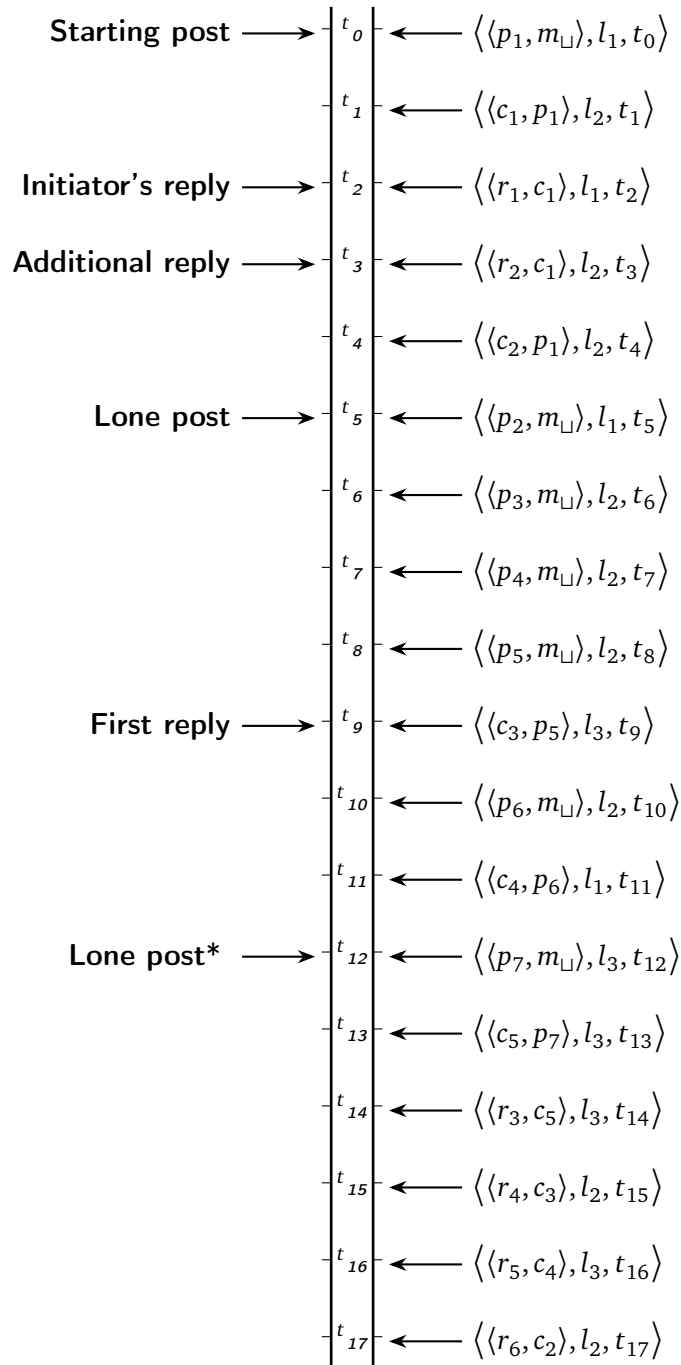


FIGURE 4.8: Timeline of e-tivities in the illustrated scenario, showing various types of interactions. Each communicative e-tivity falls into one of five categories: starting posts, lone posts, first replies, additional replies, and initiator's replies. \*Note that post  $\langle \langle p_7, m_{\perp} \rangle, l_3, t_{12} \rangle$  is still a **lone post**, since its only comment  $\langle \langle c_5, p_7 \rangle, l_3, t_{13} \rangle$  and reply  $\langle \langle r_5, c_4 \rangle, l_3, t_{16} \rangle$  are both from the initiator,  $l_3$ .

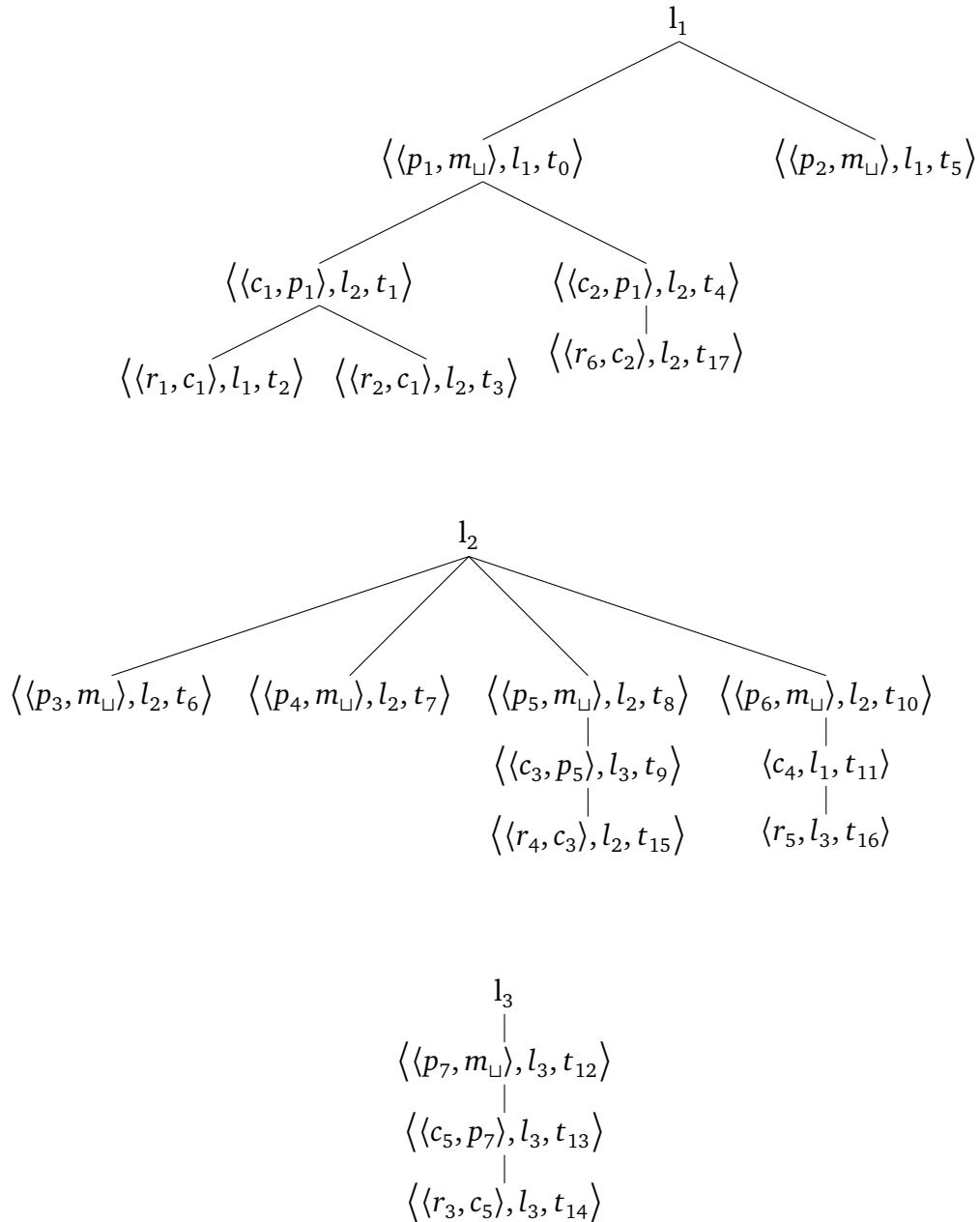


FIGURE 4.9: Alternative representation of the illustrated scenario, showing contextual relationship between posts, comments and replies. Here, a learner  $l_i$  makes a post  $p_j$  (at depth=1 in the tree), which in turn may raise a comment  $c_k$  from learner  $l_x$  (at depth=2). Learner  $l_y$  then makes a reply  $r_m$  to a comment (at depth=3).

generalisable. In particular, we can associate each of the utterances typed in by learners (and that are shown in Figure 4.5) to labels  $p_i, c_j, r_k$ . This association is made according to the role each of these utterances play in the dialogues, evidenced in the structure of the tree in Figure 4.6, where zero-order replies are all the nodes of depth=1 (at level one in the tree); first-order replies are all the nodes of depth=2 (at level two); and second-order replies are all the nodes of depth=3 (at level three).

Applying this generic representation of communicative activities  $a \in A_C$  through labels  $p_i, c_j, r_k$  to replace learners' utterances in the timeline of Figure 4.7 results into the timeline of Figure 4.8. Similarly, applying these label mappings onto the colourful tree shown in Figure 4.6 results into the tree shown in Figure 4.9.

The generic representation provided by the model is able to capture various contextual relationships amongst learner activities, namely, who responds to what, when are responses made, what conversations are non-starters, which ones spark a lot of interaction, and so on. Having this representation rooted on the learner helps to characterise said learner too, and thus can inform feature engineering for categorising learners based on their engagement in these kinds of environments, in the way shown in Sections 5.4 and 6.4 about engagement in FutureLearn MOOCs and in PeerWise, respectively.

## 4.4 Non-communicative e-tivities

At the start of this chapter, I mentioned that observing electronically-captured learning activities with a very fine granularity lends itself to interpret them as if they were atomic: either completed or not attempted.

For example, on posting a comment, it is typically irrelevant when the learner started writing a comment, which is why I have been referring about time in communicative e-tivities as timestamps  $t_k$ . However, for some other learning activities, and typically for the non-communicative kind, (such as for undertaking quizzes or a longer type of activity) it might be of interest to capture when the activity started and finished, or whether the activity is still ongoing or abandoned as unfinished, as opposed to other ones where this information is unnecessary or not captured. In other words, independent of the platform, some activities will be considered as atomic, and others as occurring over an interval of time. Therefore, to include a generalisable notion of time within the platform-agnostic model of learner interactions, time can be considered either as a snapshot (a moment, a timestamp) or an interval (a period).

### 4.4.1 Time beyond a timestamp: Intervals

More formally, let us define the following sets:

**Definition 4.17** (Timestamps). A *timestamp*  $t_k$  is a number representing a moment in time as captured electronically.  $T$  is the set of timestamps with a total order  $<$ .

**Definition 4.18** (Intervals). An *interval*, or period, is defined as a tuple  $\langle t_{start}, t_{end} \rangle \in D$  such that  $t_{start} \in T$  is a timestamp defining the start of a period, and  $t_{end} \in (T \cup \{t_\infty\})$  is either a timestamp defining the end of the period, or  $t_\infty$ , the supreme of  $T$ , satisfying  $\forall t_i \in T, t_i < t_\infty$ .

Let  $I$  be the set of intervals:

$$D \subseteq T \times (T \cup \{t_\infty\}) \quad (4.16)$$

$I$  is governed by Allen's interval algebra, which was presented in Section 2.6 of Chapter 2. This means that temporal relations between periods can be expressed formally and such be used for automated reasoning.

**Definition 4.19** (Unfinished event). An *unfinished* event (either because it has been abandoned or it is still *ongoing*) is an e-tivity with an open interval, i.e. an event which does not have (yet, at the time of the observations) an end-of-period timestamp, so it is regarded to have a finishing time of  $t_\infty$ , the supreme of  $T$ .

### 4.4.2 A practical scenario

Let us consider another scenario, involving the same learners as in the hypothetical scenario previously presented (in Section 4.3.1), Ana, Bob and Cam (later referred to as  $l_1, l_2, l_3$ ). In this scenario, these learners are engaging in non-communicative activities  $w_1, w_2, w_3$ , which are not atomic activities, in contrast with the types of activities they engaged with in the previous scenario. That is, a learner  $l_j$  will work on activities  $w_i$  over an interval (or period) of time  $d_k$ , each of which, with a start and possibly an end, with  $j \in L, w \in A, k \in \mathbb{N}$ .

This scenario is represented in Figure 4.10, with eight triples  $\langle w_i, l_j, d_k \rangle$  being the non-communicative e-tivities recorded in this environment. These e-tivities are represented as segments whose length are dependent on the duration of the interval (as indicated in the timestamps axis), with circles as delimiters to the left and right of the segment. An empty circle to the left indicates the start of an e-tivity whilst a full circle

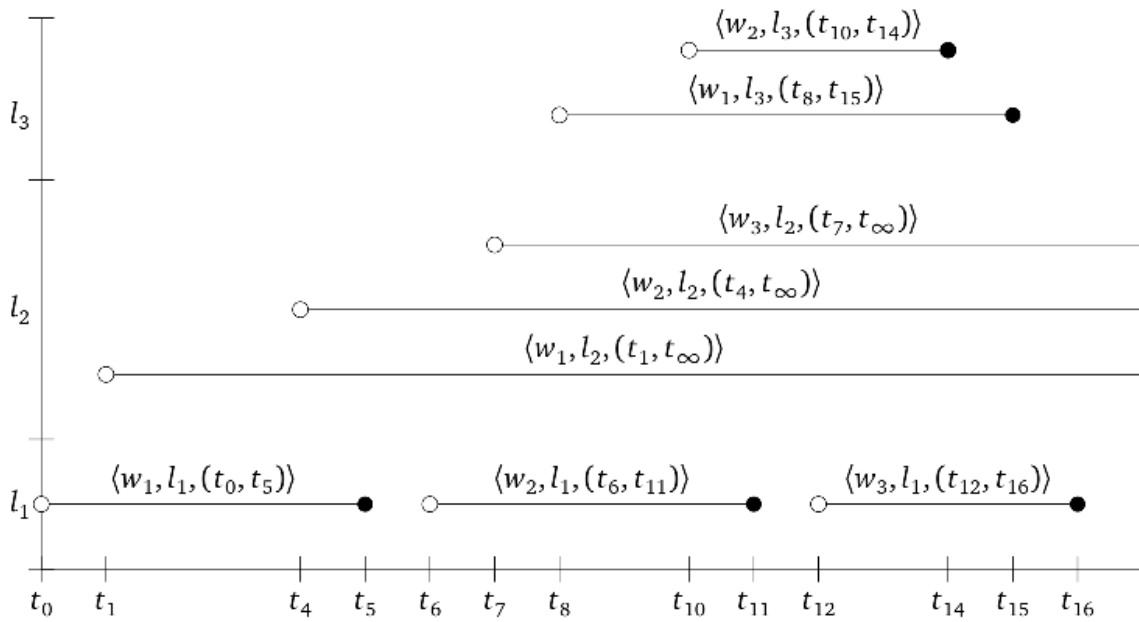


FIGURE 4.10: Timeline of non-communicative e-tivities in a second scenario, showing various types of behaviour amongst learners  $l_1, l_2, l_3$  working in activities  $w_1, w_2, w_3$  over time. Starts and ends of activity are shown by empty and full circles. The absence of a full circle indicates an ongoing or abandoned activity (so a finishing time of  $t_\infty$  is assigned.)

to the right of the segments indicates the end. A line with no terminators represents that the activity is still ongoing (or was abandoned.)

The graphical representation of the e-tivities as interval events allows visualising how these three learners have different patterns in approaching their activities. Ana comes across as somewhat methodical, as she starts and finishes each activity before embarking on the next one in the list. Bob, by contrast, seems to sample all the activities in quick succession but does not appear to finish them, whilst Cam did a couple of them in parallel, or perhaps revisited the first activity after having completed his second one.

In general, any non-communicative e-tivity can be described by its relation with the one commencing immediately after its start (by the same learner). From the set of thirteen relationships in Allen's interval algebra, shown in Table 2.7, only *before* ( $\prec$ ), *overlaps* (o), and *during* (d) are detailed in this model<sup>5</sup>. More formally:

<sup>5</sup>Inverse relationships are not detailed in this model since they do not add expressiveness to it. Also, relationships in which two intervals have the same start or ending, as described in Allen's algebra (i.e. *equal*, *meets*, *start*, *finishes*) are very unlikely to occur in this context, as it based on timestamps for starting and finishing e-tivities. Though it could be regarded as unfeasible that two separate activities by the same learner have the same timestamp, in practice these may occur due to representation errors.

**Definition 4.20** (Preceding e-tivity). An e-tivity  $\langle w, l, (t_s, t_e) \rangle$  is said to be *preceding* if there is an e-tivity  $\langle w', l, (t'_s, t'_e) \rangle$ , (by the same learner), such that  $t_s < t'_s$  and  $t_e < t'_e$ , therefore satisfying:

$$\langle w, l, (t_s, t_e) \rangle \prec \langle w', l, (t'_s, t'_e) \rangle \quad (4.17)$$

and there is no other e-tivity  $\langle w'', l, (t''_s, t''_e) \rangle$  such that  $t_s < t''_s < t'_s$ .

The operator  $\prec$  in Equation 4.17 is the *before* operator (also called *precede*) in Allen's interval algebra. For example, the e-tivity  $\langle w_1, l_1, (t_0, t_5) \rangle$  in Figure 4.10 is “preceding” because  $\langle w_1, l_1, (t_0, t_5) \rangle \prec \langle w_2, l_1, (t_6, t_{11}) \rangle$ .

**Definition 4.21** (Overlapping e-tivity). An e-tivity  $\langle w, l, (t_s, t_e) \rangle$  is said to be *overlapping* if there is an e-tivity  $\langle w', l, (t'_s, t'_e) \rangle$ , (by the same learner), such that  $t_s < t'_s$  and  $t_e < t'_e$ , therefore satisfying:

$$\langle w, l, (t_s, t_e) \rangle \circ \langle w', l, (t'_s, t'_e) \rangle \quad (4.18)$$

and there is no other e-tivity  $\langle w'', l, (t''_s, t''_e) \rangle$  such that  $t_s < t''_s < t'_s$ .

The operator  $\circ$  in equation 4.18 is the *overlaps* operator in Allen's interval algebra, shown in Table 2.7.

Not included in Allen's algebra, but evidently important in this context, is the notion of *unfinished* e-tivities, such as all of those by  $l_2$ . However, note that, somewhat counter-intuitively, according to Definition 4.21, unfinished e-tivities are not “overlapping”. This is intended, as the overlap due to a learner having abandoned an activity is different from the overlap due to a learner having returned to a previously unfinished activity after completing its subsequent one.

**Definition 4.22** (During e-tivity). An e-tivity  $\langle w, l, (t_s, t_e) \rangle$  is said to be *during* if there is an e-tivity  $\langle w', l, (t'_s, t'_e) \rangle$ , (by the same learner), such that  $t_s < t'_s$  and  $t_e > t'_e$ , therefore satisfying:

$$\langle w, l, (t_s, t_e) \rangle d \langle w', l, (t'_s, t'_e) \rangle \quad (4.19)$$

and there is no other e-tivity  $\langle w'', l, (t''_s, t''_e) \rangle$  such that  $t_s < t''_s < t'_s$ .

In the above example, the e-tivity  $\langle w_1, l_3, (t_8, t_{15}) \rangle$  is during, because  $\langle w_1, l_3, (t_8, t_{15}) \rangle d \langle w_2, l_3, (t_{10}, t_{14}) \rangle$  which is in turn due to  $t_8 < t_{10}$  and  $t_{15} > t_{14}$ .

Ultimately, the descriptions of a given learner's e-tivities can be used to characterise their engagement, such as we have seen in the scenario above. These patterns, and others, are likely to be part of a great diversity of learner engagement patterns and therefore this model can prove useful to understand the digital traces of interaction we must examine when studying such phenomena.

### 4.4.3 Putting it all together

Until now we have considered communicative activities and non-communicative e-tivities separately, but in reality they coexist, and learners may dip in and out their individual engagement with the learning materials and turn to their peers for support or for community building. As a result, the timeline of e-tivities is interspersed with e-tivities of the form  $\langle a, l, t \rangle$  and  $\langle w, l, d \rangle$ , where  $a$  is a communicative activity,  $w$  a non-communicative activity,  $l$  a learner,  $t$  a timestamp and  $d$  a period or interval. Therefore, for the scenario currently under consideration, in addition to the timestamps shown in Figure 4.10, there are five of timestamps associated to communicative e-tivities, one for each kind:

**SP**  $\langle \langle p_1, m_{\square} \rangle, l_1, t_2 \rangle$  : Ana ( $l_1$ ) makes the first post ( $p_1$ ) at time  $t_2$

**LP**  $\langle \langle p_2, m_{\square} \rangle, l_1, t_3 \rangle$  : Ana ( $l_1$ ) makes a second post ( $p_2$ ) at time  $t_3$

**FR**  $\langle \langle c_1, p_1 \rangle, l_2, t_9 \rangle$  : Bob ( $l_2$ ) comments on Ana's  $p_1$  with  $c_1$  at time  $t_9$

**IR**  $\langle \langle r_1, c_1 \rangle, l_1, t_{13} \rangle$  : Ana ( $l_1$ ) replies to Bob's  $c_1$  with  $r_1$  at time  $t_{13}$

**AR**  $\langle \langle r_2, r_1 \rangle, l_3, t_{17} \rangle$  : Cam ( $l_3$ ) comments on Ana's reply  $r_1$  with reply  $r_2$  at time  $t_{17}$

These five communicative activities can be represented in two different ways, as shown in Figures 4.8 and 4.9 for the previous scenario. This results on the forest shown in Figure 4.11 and the timeline in Figure 4.12.

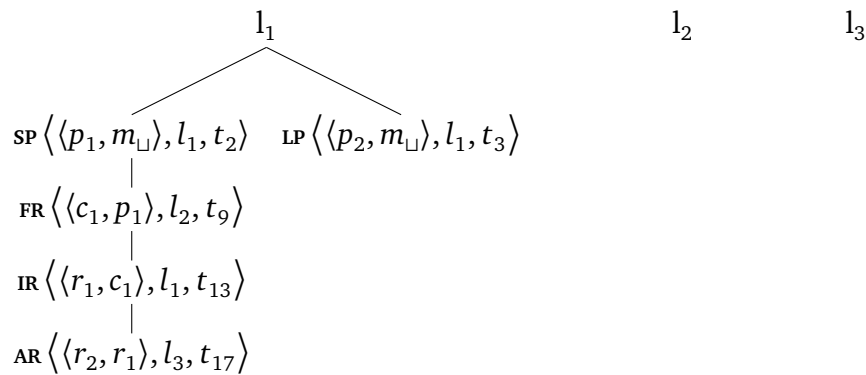


FIGURE 4.11: Forest representation of the communicative activities in the second scenario, showing contextual relationships between zero- first- and second-order replies. Here, learner  $l_1$  makes posts  $p_1$  and  $p_2$  (at depth=1), followed by a comment  $c_1$  on  $p_1$  from  $l_2$  (at depth=2), which in turn is replied by the initiator,  $l_1$  (at depth=3) with reply  $r_1$  which is then replied to by learner  $l_3$  with  $r_2$ .

These contextual relationships can also be shown alongside the non-communicative e-tivities of Figure 4.10, resulting on the timelines shown in Figures 4.13 and 4.14.

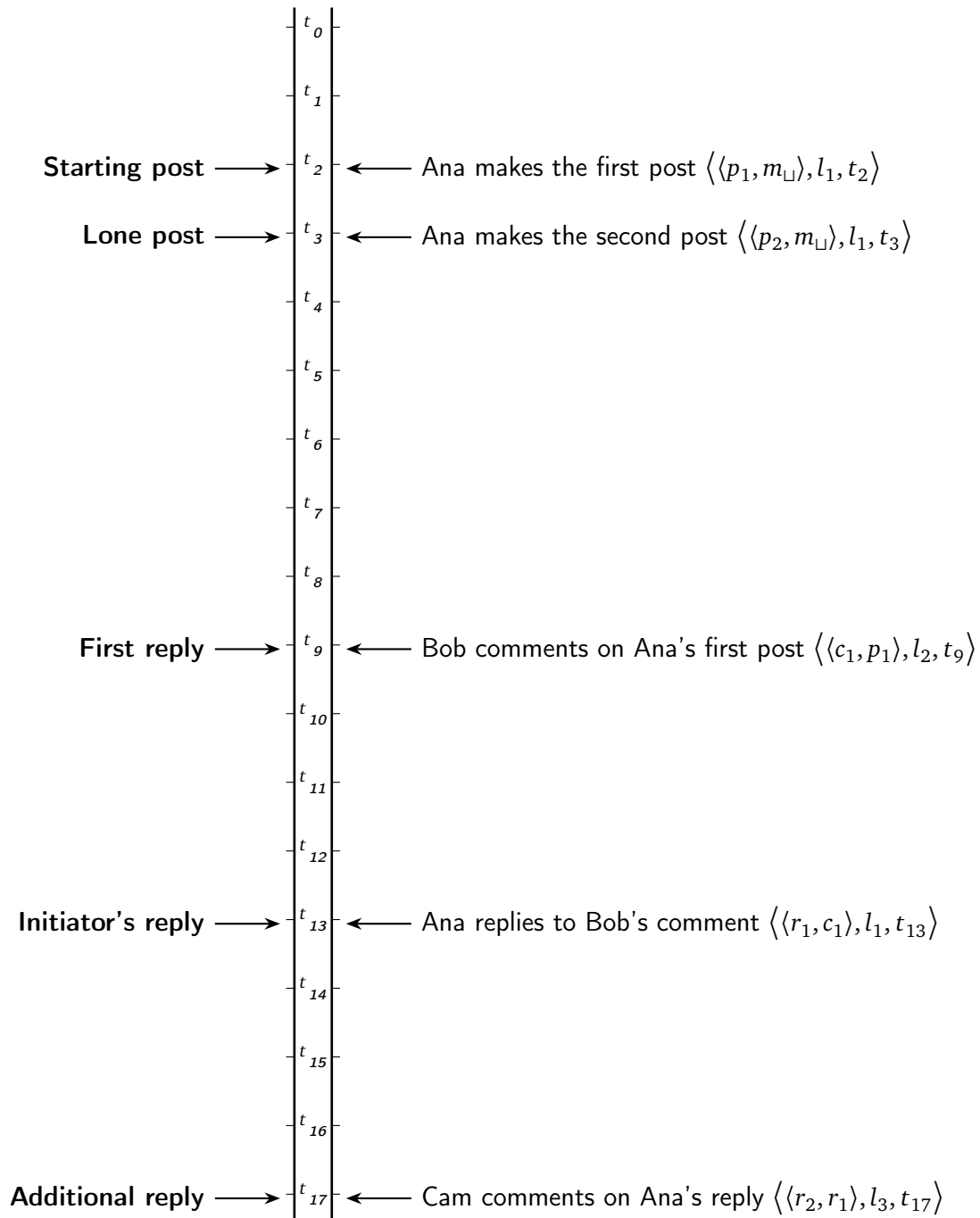


FIGURE 4.12: Timeline of the communicative activities of the scenario presented in Figure 4.11.



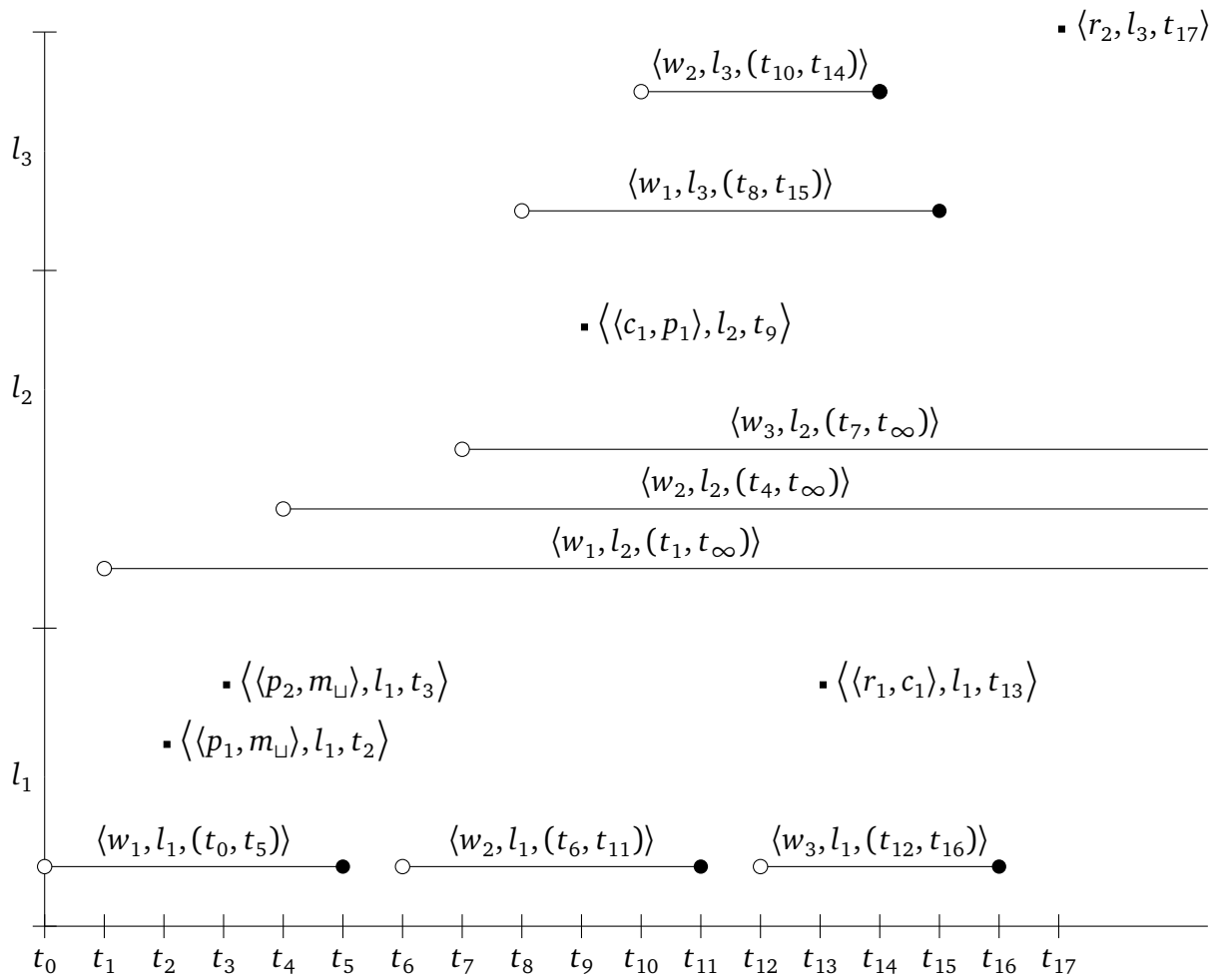


FIGURE 4.13: Timeline of communicative and non-communicative e-tivities in the second scenario, involving learners  $l_1, l_2, l_3$ . Circles indicate the starts and ends of a non-communicative activity. A square indicates a communicative activity.

## 4.5 Limitations of the model

The proposed platform-agnostic model of learner engagement within peer-supported digital environments is expressive and informative, though it presents some limitations. Earlier in the chapter I explained that a non-communicative activity may become communicative by virtue of it sparking conversations, even if it happens over a period. So far in this model this is not well represented, and there are two key reasons for it: Firstly, the model considers intentionality as a key differentiator between communicative and non-communicative activities, which is why a post that does not generate a response is a communicative activity (a “lone post”, under the categorisation).

The second limitation is around an important assumption of the model: that all communicative activities are atomic. This suggests that it may not model adequately

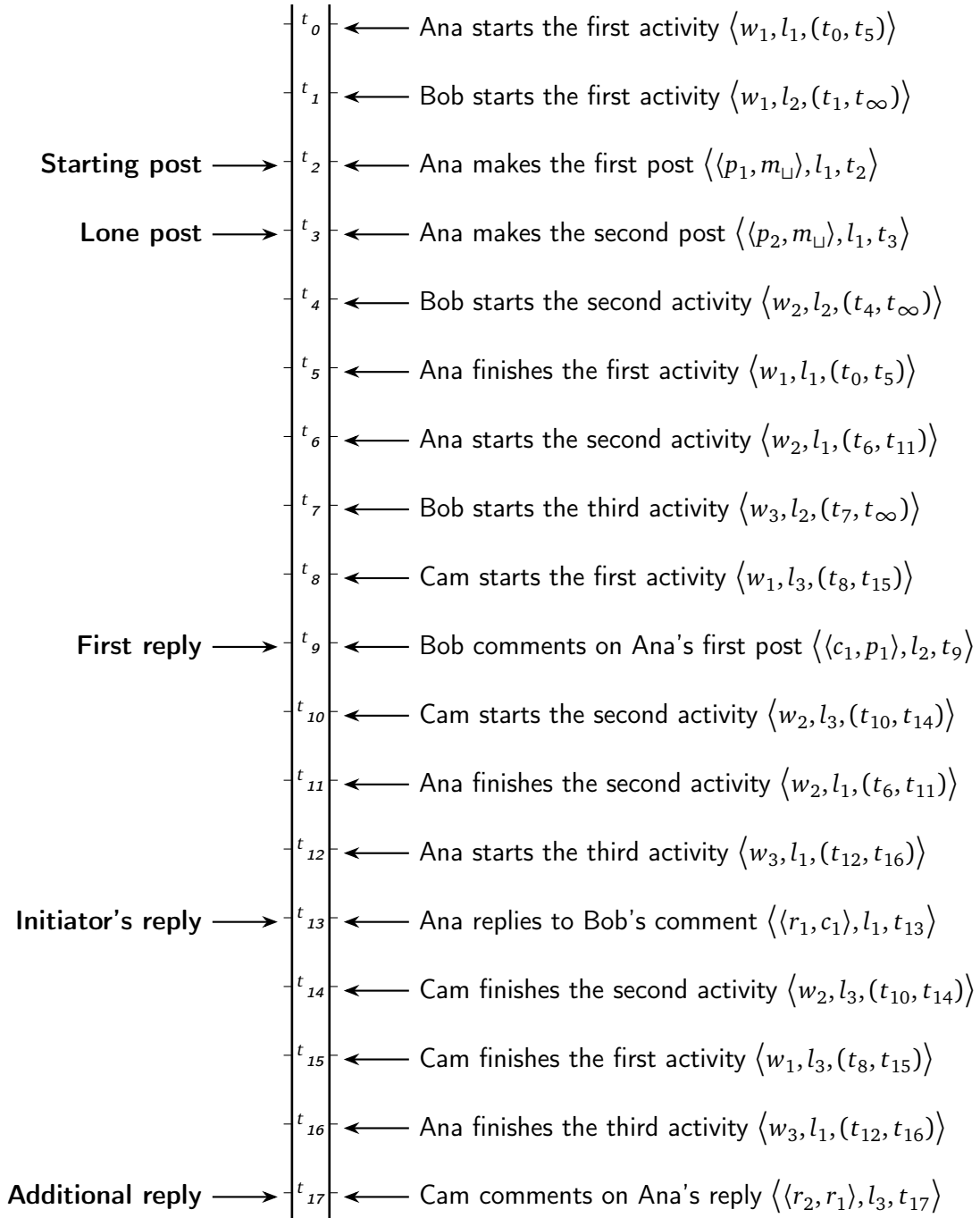


FIGURE 4.14: Timeline of the non-communicative e-tivities shown in Figure 4.10, interspersed with the communicative activities of the scenario.

some peer-supported learning platforms that support communicative activities that occur during an interval, where the start and end of the event is captured. For example, communication apps typically allow others to see whether one user have started to type, stopped typing and continue typing, even before the comment is posted, as in Figure 4.15.

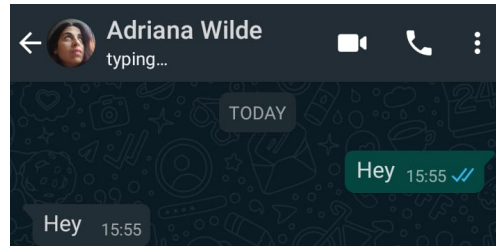


FIGURE 4.15: Screenshot in the Whatsapp messaging app whilst someone is typing: in this case the communicative activity (sending someone a message) is not atomic, not is occurring over a period of time, with the start some time prior the screenshot, and it is still ongoing. The person who with this exchange is taking place is able to see the status of this ongoing activity, in this case, “typing...”.

To illustrate another aspect of this second limitation, consider a learner intending to make some content with the purpose of communicating it with others. Even though the intent is for communication, we could conceptualise that there are in fact two different activities here: one of engagement with content (its creation, which happens over a period, which may or not be recorded as such by the platform), and a separate one of communicating it with the peers (“posting” it, which happens instantaneously). Therefore, even though the model considers communicative activities to be atomic, it can still be used to model engagement within platforms that record communicative activities as intervals, with the timestamp associated to the end of the interval activity now been assigned to be the timestamp of the atomic part of the communicative activity.

Another factor, not modelled, is visibility of the activities. It is possible, in particular for interval activities, that their visibility could have an impact on the communicative activities to be produced by others. Arguably, that in itself communicates to others that the user is engaging in the writing activity (considering the earlier discussion, showcased through Figure 4.15), which in turn may have an impact on others.

This was not addressed in this model as the design priority was on simplicity and generalisation over a wide variety of peer-learning environments. Adapting the model to include visibility of the activities as a variable is out of the scope of this thesis. Further, an alternative model which offers a good fit to that situation may suffer from “overfitting”, and therefore being less general and applicable to several other platforms which do not present such affordances.

## 4.6 Summary of the model

For ease of reference, here are all the definitions presented earlier which constitute the platform-agnostic model for learner engagement within peer-supported digital environments:

**Definition 4.1:** The set  $A$  is the set of all fine-granularity *activities*  $a_i$  that can be captured in a given peer-supported digital environment, such that  $a_i$  is either a monuple  $w_i \in A_N$  or a pair  $\langle c_i, m_j \rangle \in A_C$ . Therefore:

$$A = A_C \cup A_N$$

$$A_C \cap A_N = \emptyset$$

**Definition 4.2:** The set  $A_N \subseteq A$  is the set of *non communicative activities*  $w_i$  such that  $w_i$  is assumed to be some work undertaken within a given peer-supported digital environment over a period of time (i.e. not primarily intended for communication).

$$A_N = \{w_1, w_2, \dots, w_n\}$$

**Definition 4.3:** The set  $A_C \subseteq A$  is the set of *communicative activities*  $\langle c, m \rangle$  where  $c \in M$  is a communication in response to a message  $m \in M_0$ .  $M$  is the set of messages, and  $M_0$  is the set of messages that includes the distinguished member  $m_{\perp}$ , the *null* message (not to be confused with an empty message).

$$A_C = \{\langle c, m \rangle \mid c \in M \wedge m \in M_0\} \quad (4.20)$$

Note that  $\forall i, j : \langle c_i, m_j \rangle$ , these are instantaneous, intentional communications.

**Definition 4.4:** An electronically-captured activities, *e-tivity* is defined as a triple  $\langle a, l, t \rangle \in E \mid a \in A, l \in L, t \in T \cup D$ , to indicate an activity  $a$  ('what?') performed by a learner  $l$  ('who?') at a time  $t$  (either a timestamp or a period, 'when?'). The set  $E$  is the set of e-tivities comprising all of the activities  $a_i \in A$  undertaken by each learner  $l_j \in L$  at a time  $t_k \in T$  or over a period of time  $d_k$ , such that  $\forall t_i, t_j \in T : \text{if } i < j \text{ then } t_i < t_j$ .

$$E = \{\langle a, l, t \rangle \mid a \in A, l \in L, t \in T \cup D\}$$

**Definition 4.5:** There are two types of e-tivities:  $E = E_C \cup E_N$  and  $E_C \cap E_N = \emptyset$ .

**Definition 4.6:**  $P_{E_C} = \{E_{SP}, E_{LP}, E_{FR}, E_{IR}, E_{AR}\}$  is the partition of  $E_C$  satisfying:

$$E_C = E_{SP} \cup E_{LP} \cup E_{FR} \cup E_{IR} \cup E_{AR}$$

$$\begin{array}{llll} E_{SP} \cap E_{LP} = \emptyset & & & \\ E_{SP} \cap E_{FR} = \emptyset & E_{LP} \cap E_{FR} = \emptyset & E_{FR} \cap E_{IR} = \emptyset & E_{IR} \cap E_{LP} = \emptyset \\ E_{SP} \cap E_{IR} = \emptyset & E_{LP} \cap E_{IR} = \emptyset & E_{FR} \cap E_{AR} = \emptyset & \\ E_{SP} \cap E_{AR} = \emptyset & E_{LP} \cap E_{AR} = \emptyset & & \end{array}$$

**Definition 4.7:** The *response-to* function  $F : E_C \rightarrow E_C$  maps a communicative e-tivity  $e$  to another one  $f$ , such that  $f$  is a direct response to  $e$ , and no other e-tivity. In other words,  $F(f) = e$ , or, more precisely:

$$F \subseteq \{\langle\langle a, b \rangle, l, t \rangle, \langle\langle b, c \rangle, l', t' \rangle \mid a, b, c \in M \wedge l, l' \in L \wedge t, t' \in T\}$$

The transitive closure of  $F$ ,  $F^+$  relates a communicative e-tivity to all of its ancestors.

**Definition 4.8:** The set of (*starting posts*) is:

$$E_{SP} = \{e = \langle a, l, t \rangle \mid \exists f = \langle a', l', t' \rangle \in E_C \wedge F(f) = e \wedge l \neq l'\}$$

**Definition 4.9:** The set of (*first replies*) is:

$$E_{FR} = \{f = \langle a, l, t \rangle \mid \exists e = \langle a', l', t' \rangle \in E_{SP} \wedge F(f) = e \wedge l \neq l'\}$$

**Definition 4.10:** The set of (*initiators' replies*) is:

$$E_{IR} = \{\langle a, l, t \rangle \in E_C \mid \exists \langle a', l', t' \rangle \in E_{SP} \wedge \exists k \in \mathbb{N}, k > 1 \wedge F^k(\langle a, l, t \rangle) = \langle a', l', t' \rangle\}$$

**Definition 4.11:** The set of (*additional replies*) is:

$$E_{AR} = \{\langle a, l, t \rangle \in E_C \mid \exists \langle a', l', t' \rangle \in E_{SP} \wedge (\exists k \in \mathbb{N}, k > 1) [F^k(\langle a, l, t \rangle) = \langle a', l', t' \rangle] \wedge l \neq l'\}$$

**Definition 4.12:** The *discussion thread* function  $DT : E_C \rightarrow E_C^*$  maps a communicative e-tivity  $e_k$  with the sequence of communicative e-tivities  $e_i$ , such that each  $e_i$  is a response to another e-tivity in the sequence. This sequence includes  $e_k$  and the posts to which  $e_k$  is a direct or indirect response to (upto and including the post  $e_0$  from which the thread of responses is originated).

$$DT(e_k) = \left\{ \langle e_0, e_1, \dots, e_k \rangle \mid F(e_0) = m_{\perp} \wedge \left[ \bigwedge_{i=1}^k F^i(e_i) = e_0 \right] \right\}$$

where  $m_{\perp}$  is the distinguished member of  $M^*$ , the *null* message.

**Definition 4.13:** The set of *lone posts* is:

$$E_{LP} = \{\langle a, l, t \rangle \mid \nexists \langle a', l', t' \rangle [DT(\langle a, l, t \rangle) = \langle a', l', t' \rangle \wedge \langle a, l, t \rangle \notin DT(\langle a', l', t' \rangle)] \wedge l \neq l'\}$$

**Definition 4.14:** The set of *zero-order replies* (or posts):

$$P = \{\langle \langle p_j, m_{\perp} \rangle, l, t \rangle \mid p_j \in M, m_{\perp} \in M^*, l \in L, t \in T\}$$

where  $m_{\perp}$  is the distinguished member of  $M^*$ , the *null* message.

**Definition 4.15:** The set of *first-order replies* (or comments) is:

$$C = \{\langle \langle c_n, p_j \rangle, l_i, t_k \rangle \mid F(c_n) = p_j \wedge p_j \in P\}$$

**Definition 4.16:** The set of *second-order replies* is:

$$\langle \langle r_n, a \rangle, l_i, t_k \rangle \ni ((a = c_k \wedge c_k \in C) \vee (a = r_m \wedge r_m \in R)) \wedge F(r_n) = a \wedge m, n = 1..|R| \wedge m < n$$

**Definition 4.17:**  $T$  is the set of recorded *timestamps*.

$t_{\infty} \notin T$  is the supreme of  $T : t_{\infty} > t_i \quad \forall t_i \in T$ .

**Definition 4.18:**  $D \subseteq T \times (T \cup \{t_{\infty}\})$  is the set of periods (or intervals).

**Definition 4.19:** An *unfinished* event (either because it has been abandoned or it is still *ongoing*) is an e-tivity with an open interval, i.e. an event which does not have (yet, at the time of the observations) an end-of-period timestamp, so it is regarded to have a finishing time of  $t_{\infty}$ , the supreme of  $T$ .

**Definition 4.20:** An e-tivity  $\langle w, l, (t_s, t_e) \rangle$  is said to be *preceding* if there is an e-tivity  $\langle w', l, (t'_s, t'_e) \rangle$ , (by the same learner), such that  $t_s < t'_s$  and  $t_e < t'_e$ , therefore satisfying:

$$\langle w, l, (t_s, t_e) \rangle < \langle w', l, (t'_s, t'_e) \rangle$$

and there is no other e-tivity  $\langle w'', l, (t''_s, t''_e) \rangle$  such that  $t_s < t''_s < t'_s$ .

**Definition 4.21:** An e-tivity  $\langle w, l, (t_s, t_e) \rangle$  is said to be *overlapping* if there is an e-tivity  $\langle w', l, (t'_s, t'_e) \rangle$ , (by the same learner), such that  $t_s < t'_s$  and  $t_e < t'_e$ , therefore satisfying:

$$\langle w, l, (t_s, t_e) \rangle \circ \langle w', l, (t'_s, t'_e) \rangle$$

and there is no other e-tivity  $\langle w'', l, (t''_s, t''_e) \rangle$  such that  $t_s < t''_s < t'_s$ .

**Definition 4.22:** An e-tivity  $\langle w, l, (t_s, t_e) \rangle$  is said to be *during* if there is an e-tivity  $\langle w', l, (t'_s, t'_e) \rangle$ , (by the same learner), such that  $t_s < t'_s$  and  $t_e > t'_e$ , therefore satisfying:

$$\langle w, l, (t_s, t_e) \rangle \text{ d } \langle w', l, (t'_s, t'_e) \rangle$$

and there is no other e-tivity  $\langle w'', l, (t''_s, t''_e) \rangle$  such that  $t_s < t''_s < t'_s$ .

## 4.7 Conclusion of this chapter

The set of definitions listed in Section 4.6 constitute collectively a novel platform-agnostic model of learner engagement within peer-supported digital environments. This model provides a common language to express relationships captured in the logged interactions within these environments, such as temporal relationships, engagement with content, connections with peers, and others. In particular, both the dialogic analysis as well as the temporal analysis of learners engagement can offer valuable insights on categories of learners.

This model provides one possible answer to the research question **RQ1** (*How can learner engagement be meaningfully compared across peer-supported digital environments?*) posed in the introduction of this thesis, in Chapter 1. However, for this assertion to hold rigorously, it is important for the defined model to be validated. A strategy for such validation would consist in procuring learner-interaction data from two very different selected platforms, applying the model to inform feature engineering to such data, and obtaining the same feature sets associated to each of the datasets under comparison. The resulting analysis, if providing meaningful insights, would be a valid comparison of peer-supported digital environments that transcends the limitations of the differences between their implementations. Further, the application of the model in informing what features are important to extract, out the given digital traces of learner interactions, is likely to offer insights contributing towards an understanding of how students learn.

Chapters 5 and 6 present the validation of this model following the outlined strategy, in two different peer-supported digital environments, FutureLearn MOOCs and Peer-Wise.





## Peer-learning online within FutureLearn MOOCs

*“The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful”*

George Edward Pelham Box, FRS (b. 18 October 1919 – d. 28 March 2013)  
In Box, G. E. P.; Hunter, J. S.; Hunter, W. G. (2005), “STATISTICS FOR EXPERIMENTERS”, (2nd ed.), John Wiley & Sons.

The above quote by British statistician George Box is widely used to warn researchers about the limitations of theoretical models: whilst they might be useful in understanding underlying phenomena and main forces at play, they are invariably too simple to capture the intricate details of real applications that are subject to other forces that the model may make abstraction of. Still, its usefulness is the extent to which a particular application can be understood when studied with the generic model despite it not being a perfect fit or conforming entirely to reality (being “wrong”, as Box puts it).

The extent to which the above applies to the model presented in the previous chapter, Chapter 4, is what concerns this chapter and the next (Chapter 6). If using the model to study learner interactions in real peer-supported digital environments allows some useful insights, this model is considered to be validated. In particular, this chapter investigates **RQ2**: *What does a data-driven approach to learner interactions reveal about learning engagement within FutureLearn MOOCs?*

To answer this question, I analysed data from two of such MOOCs, provided by

the University of Southampton between 2014 and 2019. Data from seventeen runs in total<sup>1</sup>, and 271,851 enrolled learners, was studied applying the methodology described in Chapter 3 under the lens of the theoretical model formulated in Chapter 4.

This chapter is organised as follows: Section 5.1 outlines my earlier approaches to experimentation with FutureLearn MOOC data. Section 5.2 describes the datasets used in this research, followed by Section 5.3, where I explain what is observable from this data when applying a heuristic from the literature in Chapter 2. Section 5.4, then lists the features engineered from these datasets, amongst which a reduced feature set is selected as described in Section 5.5. These selected features are used in a clustering algorithm, chosen as detailed in Section 5.6, and its results presented in Section 5.7. Finally, a discussion of the findings and a conclusion for the chapter are given in Section 5.8 respectively.

## 5.1 Motivation and context

Prior to the development of the model of learner interaction from Chapter 4, I engaged in several studies using MOOC data. One of such studies was a step-centred analysis of the first run of Understanding Language (described in Appendix G). There I sought to ascertain what characteristics of a learning step, if any, led participants to complete it. I engineered fourteen features from the file `step-activity.csv` for this run of the MOOC<sup>2</sup>, together with additional information regarding the type of step. Whilst the study itself was inconclusive, it allowed me to be exposed to challenges<sup>3</sup> of data manipulation with WEKA and Tableau, as well as gaining experience with feature engineering from MOOC data.

With Cobos et al. (2017), I studied data from the second run of Portus to compare against the edX course “El Quijote” provided by the Universidad Autónoma de Madrid. This experience was an excellent exposure to heterogeneous data, and feature engineering to extract the same features from data across different platforms, such as FutureLearn and edX. The approach in that collaboration was to look only at the features that lie on the intersection of both platforms rather than trying to engineer from one those extractable in the other. Though the focus of that research was prediction of attrition

---

<sup>1</sup>Sixteen courses, with 195,465 enrolled learners in total, when excluding the third run of the Understanding Language MOOC, as explained in Chapter 3.

<sup>2</sup>The structure of this particular file is shown in Figure 3.3 and explained in section 3.2.2, alongside that all of the other files in the datasets used in this thesis.)

<sup>3</sup>Including a common gripe amongst newcomers to FutureLearn data analysis: step 1.1 being incorrectly conflated with step 1.10, when they are the first and tenth step respectively, as mentioned in section 3.2.3 and discussed during a FLAN meet-up (Wilde, Zaluska, and Millard (2015)).

(which is not the focus of this thesis), considerable effort was spent on identifying the features with the highest predictive power, and we found that number of comments is one of the two most valuable attributes for Portus, whereas for “El Quijote” it was not. One reasonable explanation was that the conversational approach in FutureLearn provided natural affordances that facilitated peer interactions in a greater way than in edX. However, it is also possible that the interactions that did occur amongst learners in “El Quijote” in particular were of less importance than non-communicative activities in terms of achieving certificate eligibility, and hence had less predictive power for this particular outcome.

In [Wilde \(2015b\)](#), I talked about my considered approach for understanding learner success in MOOCs and wanting to tend a bridge to how it is studied in F2F instruction. It proved to be a bridge built on a theoretical model of learner interactions which not only works with MOOC data (as in the research by [Chua et al. \(2017\)](#), where I drew inspiration from for part of the model) but is generalisable enough to be applied to other peer-supported learning environments, such as those facilitating face-to-face instruction as discussed in Chapter 6, which mirrors the structure of the present chapter.

## 5.2 Datasets

As mentioned in section 2.2.1, FutureLearn captures digital traces of learner activity in all of their courses, which are shared with its course providers. In particular, the University of Southampton has produced over twenty courses on this platform, amongst which, data from two courses were made available to me following ethical approval, a Data Protection Impact assessment and data management processes<sup>4</sup>.

The first of the two courses under study is the 6-week-long MOOC titled “*Archaeology of Portus: Exploring the Lost Harbour of Ancient Rome*” (called Portus herein), and the second is the 4-week long “*Understanding Language: Learning and Teaching*”, produced in collaboration with the British Council (called Understanding Language herein). Table 5.1 shows a summary of their characteristics, for each of their offerings, or “runs” as FutureLearn denotes them. Namely, the starting dates as declared in Class Central<sup>5</sup>, the number of enrolled learners and the number of active learners (as recorded in the files `enrolments.csv` and `step-activity.csv`). This means that a

---

<sup>4</sup>Details about the ethical approval process, the Data Protection Impact Assessment (DPIA) and the Data Management Process (DMP) are given in Appendices A, B and C, respectively.

<sup>5</sup>Information about MOOCs from all over the world is aggregated and curated by Class Central. Relevant listings for these two courses are available at: <https://www.classcentral.com/course/portus-1863> and <https://www.classcentral.com/course/understanding-language-2450> (Last accessed 12 February 2021).

learner would be counted as ‘enrolled’ if their `learner_id` appeared in the corresponding `enrolments.csv` file, even if they had unenrolled (which does not cause the entry to be removed, but rather, a timestamp is added in the corresponding field `unenrolled_at` in the file). Similarly, an ‘active learner’ is counted as such if their `learner_id` appeared in the corresponding `step-activity.csv` file, even if their only activity was to visit one step and never complete it.

TABLE 5.1: Enrolled and active learners per offering of each course (run) as extracted from the datasets. (Source of starting dates per run: Class Central<sup>5</sup>).

Course	Run	Started on	Enrolled learners	Active learners
<i>Archeology of Portus: Exploring the Lost Harbour of Ancient Rome</i>	1	19 May 2014	7,773	5,076
	2	26 January 2015	8,920	4,031
	3	15 June 2015	3,252	1,554
	4	13 June 2016	5,172	2,455
	5	30 January 2017	4,266	2,249
	6	26 February 2018	1,286	967
<i>Understanding Language: Learning and Teaching</i>	1	17 November 2014	58,781	27,957
	2	4 April 2015	41,912	20,435
	3	19 October 2015	44,283	N/A
	4	4 April 2016	25,590	11,716
	5	17 October 2016	19,872	10,947
	6	24 April 2017	10,278	5,346
	7	8 January 2018	12,899	8,447
	8	11 June 2018	6,033	3,015
	9	22 October 2018	8,310	5,795
	10	29 April 2019	5,095	3,067
	11	21 October 2019	7,831	4,101
	12	27 April 2020	N/A	N/A
	13	12 October 2020	N/A	N/A

### 5.2.1 Learning design changes

There are additional observations to make on Table 5.1. The third run of the Understanding Language MOOC has 44,283 enrolled learners but the table does not provide information about active learners. The reason for this information not to be available is that, as explained in section 3.2.3, this run was excluded from subsequent data analysis due to inconsistencies that were not easily fixable in the data cleaning phase of the methodology. Other missing fields in the table are to be noted for runs 12 and 13 of the Understanding Language MOOC. Though to date there have been two further offerings of this course, only some data from the first 11 were made available to me. Further details regarding the data collection process are given in Appendices A and B.

Due to small changes in course design between runs of each MOOC, the number of steps that can be visited (and completed) by learners vary as shown in detail in Table 5.2. These adjustments to the learning design between consecutive runs of a given MOOC may have had an effect on the engagement. For example, in Portus there was a small change to the number of steps between the first and second run which perhaps had an effect on learning engagement, even though that the structure of the MOOC was very similar (there was only one less step in the last week of the course). However, between the third and fourth runs, the structural changes not only seem to be greater<sup>6</sup>, but also occurring earlier in the MOOC, with three steps fewer in week one and in week six.

TABLE 5.2: Steps per week in each run for both courses, as extracted from the datasets.

Course name	run	steps per week						num. steps
		1	2	3	4	5	6	
<i>Archaeology of Portus: Exploring the Lost Harbour of Ancient Rome</i>	1	1.1 - 1.23	2.1 - 2.21	3.1 - 3.19	4.1 - 4.18	5.1 - 5.23	6.1 - 6.17	121
	2	1.1 - 1.23	2.1 - 2.21	3.1 - 3.19	4.1 - 4.18	5.1 - 5.23	6.1 - 6.18	122
	3	1.1 - 1.23	2.1 - 2.21	3.1 - 3.19	4.1 - 4.18	5.1 - 5.23	6.1 - 6.18	122
	4	1.1 - 1.20	2.1 - 2.21	3.1 - 3.19	4.1 - 4.18	5.1 - 5.23	6.1 - 6.15	116
	5	1.1 - 1.20	2.1 - 2.21	3.1 - 3.19	4.1 - 4.18	5.1 - 5.23	6.1 - 6.15	116
	6	1.1 - 1.20	2.1 - 2.21	3.1 - 3.19	4.1 - 4.18	5.1 - 5.23	6.1 - 6.15	116
<i>Understanding Language: Learning and Teaching</i>	1	1.1 - 1.17	2.1 - 2.15	3.1 - 3.12	4.1 - 4.20			64
	2	1.1 - 1.18	2.1 - 2.16	3.1 - 3.12	4.1 - 4.20			66
	3	1.1 - 1.18	2.1 - 2.16	3.1 - 3.12	4.1 - 4.20			66
	4	1.1 - 1.18	2.1 - 2.16	3.1 - 3.12	4.1 - 4.20			66
	5	1.1 - 1.18	2.1 - 2.16	3.1 - 3.12	4.1 - 4.18	5.1 - 5.12		76
	6	1.1 - 1.18	2.1 - 2.16	3.1 - 3.12	4.1 - 4.18	5.1 - 5.12		76
	7	1.1 - 1.17	2.1 - 2.15	3.1 - 3.12	4.1 - 4.21			65
	8	1.1 - 1.16	2.1 - 2.15	3.1 - 3.12	4.1 - 4.21			64
	9	1.1 - 1.16	2.1 - 2.15	3.1 - 3.13	4.1 - 4.21			65
	10	1.1 - 1.16	2.1 - 2.15	3.1 - 3.14	4.1 - 4.21			66
	11	1.1 - 1.16	2.1 - 2.15	3.1 - 3.15	4.1 - 4.21			67

Similarly, for Understanding Language, it can also be observed in Table 5.2 that there was a very small variation in the number of steps between the first and second runs, and bigger structural changes for the fifth and six runs, all of which may have had an effect on the learning experience and the engagement. Additionally, though not observable in this table, there might have been significant changes to steps (besides additions or removals). For example, inspecting the number of videos in runs eight and nine of the Understanding Language MOOC, we can see it jumped from 31 to 34 (as shown in the *video-stats* file shape<sup>7</sup> listed in Tables D.2 and D.3 in Appendix D). Rather than three new steps in run nine, however, there is only one more (step number 3.13), which suggests the other two videos must have been added as steps that were previously of a different type.

<sup>6</sup>One of such changes was mentioned in Footnote 4, the removal of peer review assignments, as can be seen by inspecting Table D.1.

<sup>7</sup>Note that the file shape of CSV files in this dataset will always have one more row than the number of instances because the file header is included in the count.

In addition to all of these changes in course design, the FutureLearn platform itself keeps evolving and adding functionalities over the years, many of which would necessarily have an effect on learners behaviour. One such example is the functionality by which learners receive email notifications from FutureLearn when others have left comments to their posts<sup>8</sup>. All of the variations discussed above, and their potential effects, are needed to bear in mind when performing an inter-run comparison of learner engagement in the MOOC.

### 5.3 A heuristic approach to discussion analytics

The first step in analysing the MOOC data described in Section 5.2 is studying it through the lens of a heuristic from the literature that has already been applied using FutureLearn data such as, in particular, the dialogic approach used by [Chua et al. \(2017\)](#).

As per the summary of research listed on Table 2.3, this heuristic-based categorisation of learners in MOOCs was based on the identification of the types of comments created. By examining the comments learners made, in relation to their positioning in the turn-taking nature of conversations, they proposed a categorisation of learners. In their dialogic analysis, the authors distinguished five types of comments: initiating posts, lone posts, replies, initiator's replies and further replies. Based on whether learners had made comments of each of these five types or not, there would be  $2^5 = 32$  permutations of possible classes (irrespective of the number of comments per type, hence binary). However, in doing a further analysis of each combination, the authors grouped them according to their distinctive features in seven categories of social learners: 'loners'<sup>9</sup>, 'repliers', 'initiators without replying', 'initiators who respond', 'active social learners without turn-taking', 'reluctant active social learners' and 'active social learners'. Then, they applied this heuristic to comment data from the first run of the FutureLearn MOOC "Inequalities in Personal Finance: The Baby Boom Legacy", or Personal Finance in short, obtaining the counts of learners per category that are listed in Table 5.3.

Each category in this heuristic is listed below, with its definition from [Chua et al. \(2017\)](#). These definitions are complemented with the corresponding counts across the five types of communicative activities defined in the model<sup>10</sup> proposed in Chapter 4.

---

<sup>8</sup>The exact date of the release of this functionality is not easily retrievable. However, I am aware that it was sometime in early 2016, as confirmed in a personal communication from Richard Banks (FutureLearn Head of Studio) via Monty King (FutureLearn Learning Designer), on 22 February 2021.

<sup>9</sup>Given its pejorative connotations, I do not agree with the use of the term 'loners' which, in this context, merely refers to learners who had not received replies to their posts. Alas, such comments are called 'lone posts', thus any other name would have perhaps been unclear.

<sup>10</sup>Figure 4.6 is particularly helpful to look at when applying these definitions.

TABLE 5.3: Absolute counts per social learner group in the first run of the Personal Finance MOOC calculated by Chua et al. (2017). Numbers in brackets include counts for educators, as per the accepted version of the manuscript (<http://oro.open.ac.uk/57071/>)

Loners	Repliers	Initiators without replying	Initiators who respond	Active social learners without turn-taking	Reluctant active social learners	Active social learners
164 (165)	40	114	37	98	5 (6)	178 (181)

- *Loners*: “Never received replies”. In terms of the model, these learners produced comments that include non-zero lone posts (LP) and zero starting posts (SP), zero initiators’ replies (IR) and zero additional replies (AR).
- *Repliers*: “Only replied to others”. Therefore these learners would have zero SP and LP but non-zero FR or AR. In the example in Figure 4.11,  $l_2$  and  $l_3$  are such a type of learners.
- *Initiators without replying*: “Never replied to others’ posts or underneath (their) own initiating post”. These learners would then have contributed non-zero SP but zero FR, IR and AR.
- *Initiators who respond*: “Never replied to others’ posts but responded to others’ replies underneath (their) own initiating post”, i.e. non-zero SP and IR but zero FR.
- *Active social learners without turn-taking*: “Initiated posts, replied to others, but never replied under own initiating post or further replied”, i.e. non-zero SP, LP and FR but zero IR and AR.
- *Reluctant active social learners*: “Created lone posts, replied to others, further replied”. In the model, this is equivalent to having zero SP and non-zero LP, FR.
- *Active social learners*: “Initiated posts, replied to others, and engaging (sic) in turn-taking by replying under (their) own initiating post or further replying”. These learners then would have non-zero comments across all of the dialogic features (though possibly zero in either IR or AR but not both). In the example in Figure 4.11,  $l_1$  belongs to such a type of learners.

One observation to make about Table 5.3 is that there are some numbers in brackets, accounting for a discrepancy between the two versions of this publication that are available online. Specifically, the accepted version<sup>11</sup> of the manuscript, (before the ed-

<sup>11</sup>The accepted version of the manuscript is available in the Open Research Online repository by the Open University (ORO) at <http://oro.open.ac.uk/57071/>, last accessed 20 February 2021.

its resulting from recommendations from the peer-review process) included the educators in the learners' counts. In contrast, the final version<sup>12</sup> of the manuscript does not. Both are shown for reference<sup>13</sup>. The difference between the numbers in brackets and those immediately above them indicates the behaviours exhibited by the educators in the MOOC. In particular, these educators fell in the following categories: one 'loner', one 'reluctant active social learner', and three 'active social learners'.

There is a final observation about Table 5.3, which is most evident when these counts are plotted in the bar chart shown in Figure 5.1. In this MOOC, the counts for active social learners make it the largest of the categories, rather than the smallest, as the 90-9-1 rule discussed in Section 2.1.3 would predict. This is important to point out although the reason for this apparent high engagement is not addressed in Chua et al. (2017).

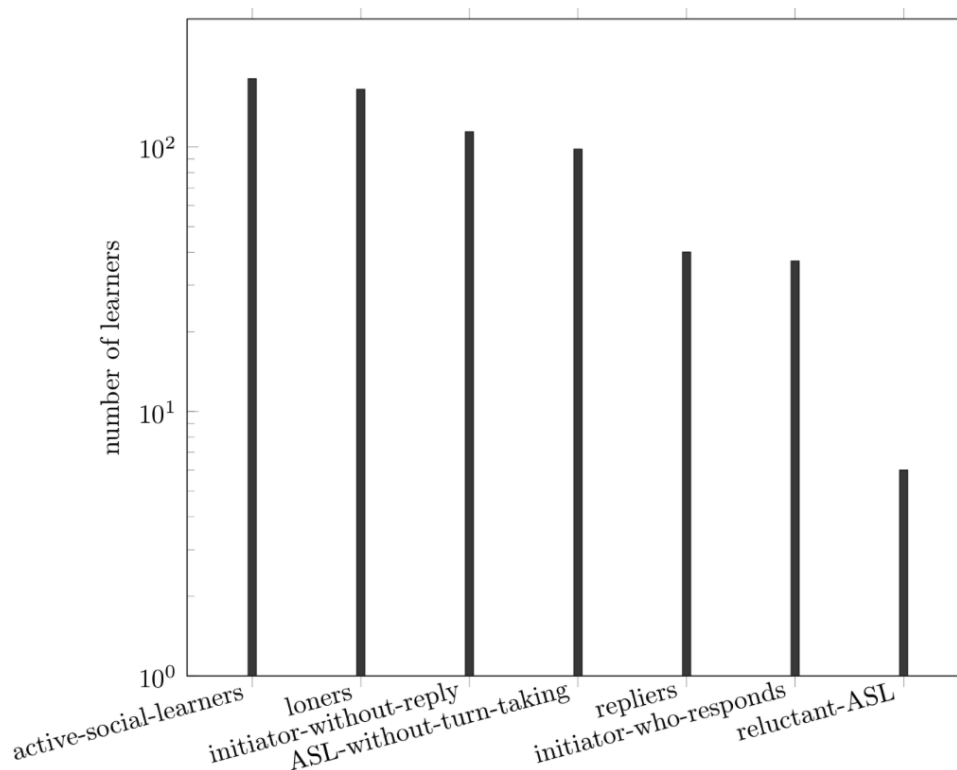


FIGURE 5.1: Distribution of social learners in the Personal Finance MOOC (from Table 5.4), in descendent order by number of learners in each category. Note that the largest category is the one comprising active social learners.

Two possible explanations are as follows. One, it is possible that the characteristics of the Personal Finance MOOC induced a much-higher-than-expected social engagement,

<sup>12</sup>The final version of the manuscript is available in CEUR at [http://ceur-ws.org/Vol-1967/FLMOOCS\\_Paper3.pdf](http://ceur-ws.org/Vol-1967/FLMOOCS_Paper3.pdf), last accessed 20 February 2021.

<sup>13</sup>In my comparisons I use the version that includes the educators because I was unable to exclude the educators from the data in the MOOCs I studied. This information would have been extractable from the `role` column in the `enrolment.csv` files had I received them without the removal of the `learner_id` column.



be it due to the learning design of this MOOC specifically or the FutureLearn platform in general. Another, more plausible explanation, is that the heuristic cannot capture the intensity of the learner engagement that the rule predicts. This is more evident in findings reported in the next section.

### 5.3.1 Applying the heuristic by Chua et al. (2017)

Applying the heuristic above defined to the datasets for Portus and Understanding Language MOOCs produces two distribution of learners which are comparable to that in Table 5.3. This table has now been extended to include the counts for each run of these MOOCs, as shown in Table 5.4. It is important to note the addition of the category ‘asocial learners’ to those by Chua et al. (2017), to include the counts of those who had not made any comments amongst the so-called active learners (as defined in Section 5.2), who had, at the very least, visited a step. It excludes, however, enrolled learners who had not. This information is not in Chua et al. (2017), as these researchers focused on social learners in their MOOC of interest.

TABLE 5.4: Absolute counts per social learner group in the categorisation by Chua et al. (2017), including counts reported by the authors for reference against experiments in this thesis. Numbers in brackets include statistics for educators.

Course name	run	Social Learners Groups found in each dataset							
		Asocial learners	Loners	Repliers	Initiators without replying	Initiators who respond	Active social learners without turn-taking	Reluctant active social learners	Active social learners
<i>Inequalities in Personal Finance: the Baby Boom Legacy</i>	1	N/A	164 (165)	40	114	37	98	5 (6)	178 (181)
<i>Archeology of Portus: Exploring the Lost Harbour of Ancient Rome</i>	1	3261	978	6	310	38	10	5	470
	2	2748	611	11	166	79	5	21	392
	3	1205	168	6	64	23	3	4	83
	4	1716	347	4	84	50	4	14	239
	5	1704	220	4	69	28	3	12	211
	6	751	89	1	18	8	0	6	96
	all	11385	2413	32	711	226	25	62	1491
<i>Understanding Language: Learning and Teaching</i>	1	16664	5990	9	1758	608	27	92	2810
	2	12865	4034	11	1256	420	8	73	1769
	4	7560	2171	8	710	286	6	47	929
	5	8519	1278	6	354	127	2	16	646
	6	4032	674	3	212	74	2	19	331
	7	6302	1241	7	238	87	5	24	544
	8	2490	384	0	125	27	0	8	182
	9	4538	646	6	170	59	4	10	363
	10	2323	389	5	120	31	2	9	189
	11	3015	655	8	141	26	1	18	238
		all	68308	17462	63	5084	1745	57	316

Another observation to make from this table is that the number of learners participating in these three MOOCs vary greatly, which can make the comparison challenging.

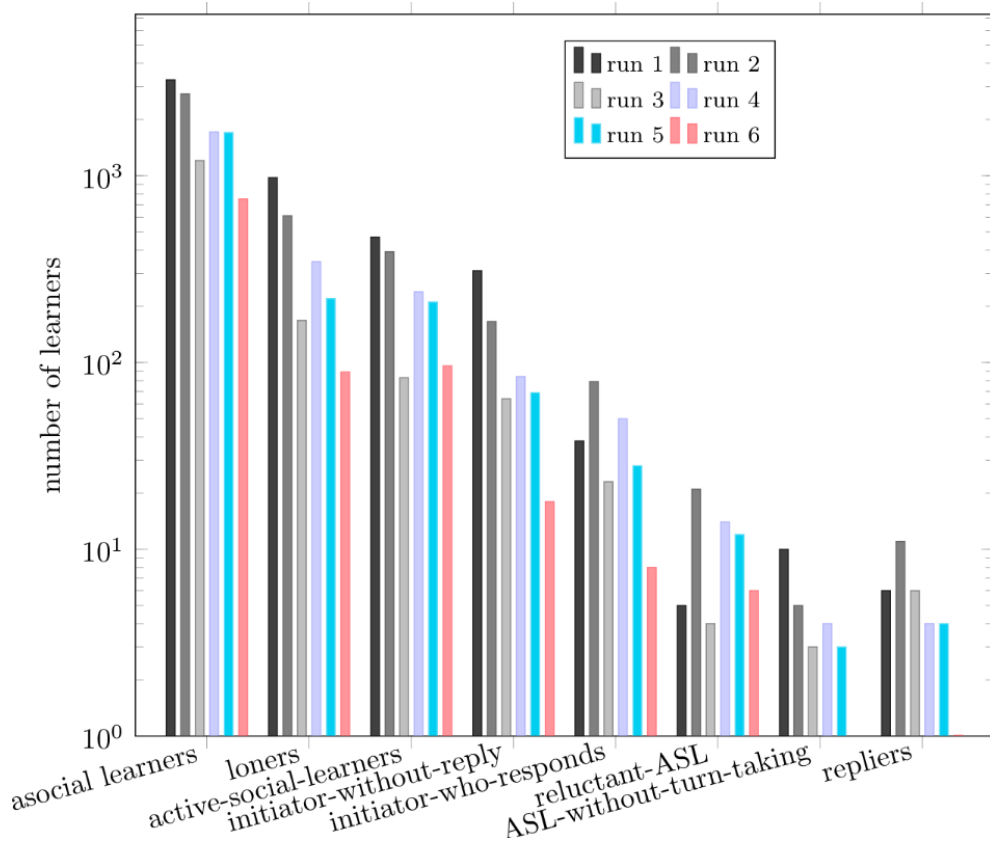


FIGURE 5.2: Distribution of learners on each of the eight categories found on applying the extended heuristic on Portus MOOCs data (shown in Table 5.4)

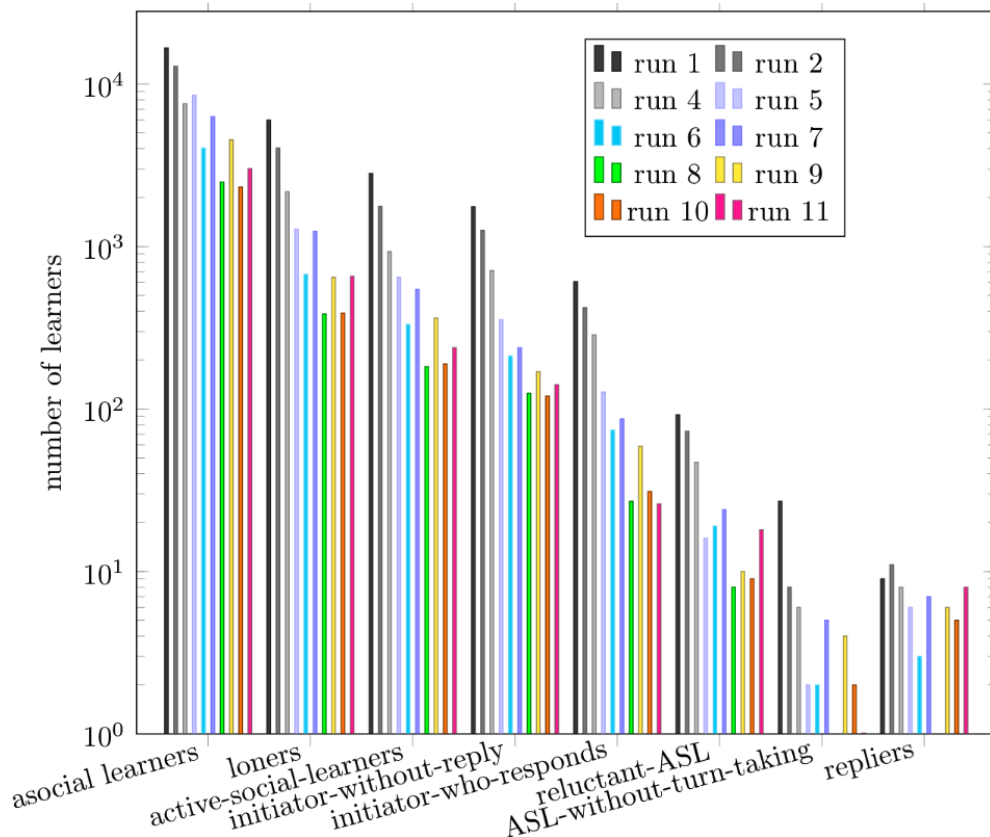


FIGURE 5.3: Distribution of learners on each of the eight categories found on applying the heuristic on Understanding Language MOOCs data (shown in Table 5.4).

Whilst in the Personal Finance MOOC there are a few hundreds of learners, the Portus MOOC has a few thousands and the Understanding Language MOOC a few tens of thousands. To facilitate the comparison, Figures 5.1, 5.2 and 5.3, use a logarithmic scale for the bar charts for each of these MOOCs. This is particularly helpful when inspecting data from the Understanding Language MOOC, where the largest category (asocial learners) dwarfs the smallest one (repliers), with 16,664 learners against nine in its first run.

Having overcome the issue about the differences in scale amongst these three MOOCs, it is easier to look for both similarities and differences in learner social behaviours across these MOOCs. For example, as noted earlier, in the Personal Finance MOOC, the largest category reported was ‘active social learners’, whereas for the other two it was ‘loners’ by far (if we exclude ‘asocial learners’ for the comparison). However ‘active social learners’ is still one of the most populated categories across all the runs of both MOOCs. The reason, as hypothesised in Section 5.3, is that the heuristic does not capture the intensity of participation as per widely-observed 90-9-1 rule. The heuristic disregards the intensity of the participation to turn the focus of the classification to the type of comments learner make, irrespective of how many of each kind they produce.

Finally, one important observation across these MOOCs is that there were learners across every category when applying this heuristic, and roughly follow similar proportions throughout the datasets, though there are some differences across MOOCs and between runs of the same MOOC.

## 5.4 Feature engineering on FutureLearn MOOC data

Having described the MOOC datasets and used a heuristic from the literature to confirm the existence of a variety of social behaviours amongst learners, in the remaining sections of this chapter I explore whether these categories can be discovered via unsupervised learning, such as by clustering. In other words, whether a data-driven approach would elicit a comparable categorisation of learning engagement in MOOCs.

Here I apply the model presented in Chapter 4. As this model characterises two types of *e-tivities*, each of part of this model was used to inform two kinds of feature sets engineered from the MOOC data, in the spirit proposed by Baker (2020) and detailed in Section 2.5.1. In particular, communicative *e-tivities* helped defining dialogic features as detailed in Section 5.4.1, whereas non-communicative *e-tivities* helped defining the interval features in Section 5.4.2. In addition to those inspired by the model, I consider other features in Sections 5.4.3 and 5.4.4.

### 5.4.1 Dialogic features

The dialogic features engineered from the model require the calculation of comment types, which is performed for each learner as per the definitions given in Section 4.3.3. The counts for each type become then the features summarised in Table 5.5.

TABLE 5.5: Dialogic features engineered (as informed by the model in Chapter 4)

Feature	Description	Equivalent comment type in Chua et al. (2017)
SP	Count of starting posts (comments created by the learner which attract replies but are not replies themselves). These are zero-order replies.	Initiating post
LP	Count of lone posts (posts created by the learner which do not attract replies from others and are replies themselves).	Lone post
FR	Count of first replies (replies to someone else's starting post). These are first-order replies.	Reply
IR	Count of initiator replies (replies to someone's reply to their own starting post). These are second-order replies.	Initiator's reply
AR	Count of additional replies (replies to a reply to a starting post created by someone else). These are also second-order replies.	Further reply

### 5.4.2 Interval features

In order to engineer these features, For each learning step visited by a learner, if completed, is given an `Event_type` value based on its timestamps relationship with respect to the timestamps of the next step visited. The relationship is determined applying Allen's algebra definitions<sup>14</sup> given in Section 2.6. If not completed, the `Event_type` of the step is 'abandoned' instead, however, if it has been completed but is the last step performed by the learner (hence there is no "next step" to compare it against), then it is assigned a `Event_type` of 'last'. The interval features engineered from the model are therefore shown in Table 5.6.

For example, consider the sequence of steps given in the `step-activity.csv` of a fictitious MOOC with three learners, shown in Table 5.7. To aid the comprehension of this example, the timestamps have been chosen to intuitively match the indices of each of the  $t_i$  timestamps shown in Figure 4.10 when intervals for non-communicative activities were first introduced. Hence,  $t_0$  becomes "2021-01-01 00:00:00 UTC",  $t_1$  becomes "2021-01-01 01:00:00 UTC", and so on.

This file is processed by constructing the sequence of steps undertaken by each learner, in this case the three sequences shown in Table 5.8. For each step in the re-

<sup>14</sup>That is, all of the direct relations defined by Allen (1983), except 'during' which is in fact is the inverse relation, i.e. if "next step" is visited and completed 'during' the current one.

TABLE 5.6: Interval features engineered (as informed by the model in Chapter 4)

Feature	Description
<b>precede</b>	Count of steps of Event_type = 'precede'
<b>overlap</b>	Count of steps of Event_type = 'overlap'
<b>during</b>	Count of steps of Event_type = 'during'
<b>abandoned</b>	Count of steps of Event_type = 'abandoned'
<b>equal</b>	Count of steps of Event_type = 'equal'
<b>begin</b>	Count of steps of Event_type = 'begin'
<b>finish</b>	Count of steps of Event_type = 'finish'
<b>meet</b>	Count of steps of Event_type = 'meet'
<b>last</b>	Count of steps of Event_type = 'last'

TABLE 5.7: Extract of the step-activity file associated to the toy example MOOC from Figure 4.10), built to illustrate the calculation of interval features

learner_id	step	...	first_visited_at	last_completed_at
learner_1	1.1.	...	2021-01-01 00:00:00 UTC	2021-01-01 05:00:00 UTC
learner_2	1.1.	...	2021-01-01 01:00:00 UTC	
learner_2	1.2.	...	2021-01-01 04:00:00 UTC	
learner_1	1.2.	...	2021-01-01 06:00:00 UTC	2021-01-01 11:00:00 UTC
learner_2	1.3.	...	2021-01-01 07:00:00 UTC	
learner_3	1.1.	...	2021-01-01 08:00:00 UTC	2021-01-01 15:00:00 UTC
learner_3	1.2.	...	2021-01-01 10:00:00 UTC	2021-01-01 14:00:00 UTC
learner_1	1.3.	...	2021-01-01 00:00:00 UTC	2021-01-01 05:00:00 UTC

sulting sequence, Allen's algebra definitions are applied with respect to the next step in the sequence when possible (with the variations described above), resulting in the event types for each of those steps, also shown in Table 5.8, and collated in the feature vectors in Table 5.9.

TABLE 5.8: Sequences of steps and their Event\_types for each learner in the toy example from Figure 4.10

learner_id	steps sequence	steps Event_types
learner_1	1.1. → 1.2. → 1.3.	precede → precede → last
learner_2	1.1. → 1.2. → 1.3.	abandoned → abandoned → abandoned
learner_3	1.1. → 1.2.	during → last

TABLE 5.9: Values of the interval features for each learner in the toy example

learner_id	precede	overlap	during	abandoned	equal	begin	finish	meet	last
learner_1	2	0	0	0	0	0	0	0	1
learner_2	0	0	0	3	0	0	0	0	0
learner_3	0	0	1	0	0	0	0	0	1

### 5.4.3 Badge features

In addition to the feature sets of Tables 5.5 and 5.6, informed by the model of learner engagement, I also included features that are possible to engineer and that would be equivalent to those directly extractable from the PeerWise dataset, as per the recommendations in 2.5.1. I pursued this with the intention of defining as many features as possible that are common to both peer-supported digital environments. Without detailing yet too much how are these features defined in PeerWise<sup>15</sup>, Table 5.10 shows three of such features. Even though FeatureLearn MOOCs do not have a gamification approach nor use badges to signal milestones, learners still reach said milestones. Therefore, a convenient feature that would facilitate a comparison across platforms is one capturing whether the learner has reached a given milestone (i.e. whether they would be eligible for a series of badges given their engagement).

TABLE 5.10: “Badge” features engineered (inspired from PeerWise badges in Chapter 6)

Feature	Description
B1	One, if the learner has posted at least a comment, zero if not.
B4	One, if the learner has posted at least a first reply, zero if not.
B5	One, if the learner has posted at least an initiators’ reply, zero if not.

### 5.4.4 Other features

In addition to the features in the three categories listed above, the following features were extracted from the `step-activity.csv` file, as listed in Table 5.11.

TABLE 5.11: Other features extracted for MOOC learners

Feature	Description
<code>steps_visited_ratio</code>	Count of steps visited by the learner over the total number of steps in the MOOC (listed in Table 5.2). Engineered from <code>step-activity.csv</code>
<code>steps_completed_ratio</code>	Count of steps completed by the learner over the total number of steps in the MOOC. Engineered from <code>step-activity.csv</code>
<code>eligible_for_certificate</code>	Calculated as ‘True’ if their <code>steps_completed_ratio</code> is greater than 0.5. ‘False’ otherwise.
<code>archetype</code>	Self-reported learning archetype from those listed in Table 2.4 (engineered from <code>archetype-survey-responses.csv</code> when available).

Finally, I engineered a final set of features by which both communicative and non-communicative activities were assigned to bins according to when in the course they were performed. Hence, there are `pre-` and `post-` features, capturing counts of each type of learner activities before and after the formal start and end of the course; there

<sup>15</sup>This is covered in Section 6.4.

are early- features, capturing learner activities in the first ten days of the course<sup>16</sup>; and finally, there are late- features, capturing learner activities after the tenth day but before the course ended.

All together, the complete set of up to 78 features<sup>17</sup> is listed as follows:

num_visited_steps	early_begin	pre_AR
num_completed_steps	early_during	pre_FR
precede	n_early_equal	pre_IR
overlap	n_early_finish	pre_LP
during	early_last	pre_SP
abandoned	early_meet	early_AR
equal	early_overlap	early_FR
finish	early_precede	early_IR
meet	late_abandoned	early_LP
last	late_begin	early_SP
num_comments	late_during	late_AR
SP	n_late_equal	late_FR
LP	n_late_finish	late_IR
FR	late_last	late_LP
IR	late_meet	late_SP
AR	late_overlap	post_AR
pre_abandoned	late_precede	post_FR
pre_begin	post_abandoned	post_IR
pre_during	post_begin	post_LP
n_pre_start_equal	post_during	post_SP
n_pre_start_finish	n_post_end_equal	B1
pre_last	n_post_end_finish	B4
pre_meet	post_last	B5
pre_overlap	post_meet	steps_visited_ratio
pre_precede	post_overlap	steps_completed_ratio
early_abandoned	post_precede	eligible_for_certificate

<sup>16</sup>Ten days were chosen as a cut-off point for early engagement, motivated by [Kizilcec and Chen \(2020\)](#), who observed that student engagement in an SMS-based mobile learning platform declined rapidly after this point.

<sup>17</sup>Any features that are zero for all the instances of a given run of a MOOC are removed from the python dataframe constructed during the feature engineering process as they would not be of any use for the analysis of that set of instances, and therefore do not appear in the resulting learner features file associated to the run of a MOOC.

## 5.5 Selecting a feature set

The feature set engineered for FutureLearn MOOC data, described in Section 5.4 above has a very high dimensionality, containing up to 78 features. This poses a problem for machine learning methods in particular, called the “curse of dimensionality”. Even though the inclusion of more features to describe a learner may characterise them better (as more information about them is available), the inclusion of each leads to instances become more separated as the search space gains an additional dimension, and therefore, organising them in groups of high similarity becomes extremely challenging (Pestov, 2013). In order to reduce the dimensionality and avoid this problem, to which clustering algorithms are particularly sensitive, there are two possible approaches, which I consider in the following sections.

### 5.5.1 Why not Principal Component Analysis?

A commonly used approach for dimensionality reduction is the application of principal component analysis (PCA), to reduce the dimensions of the data to the components that explain maximum variability (Witten et al., 2017). This approach is employed to both improve the performance of unsupervised learning algorithms, and also, typically for the convenience of representation in a two-dimensional space, e.g. figures on paper.

The method is essentially a transformation to a lower dimensional space, which relies on the calculation of orthogonal vectors, chosen successively to capture the largest variance. It does so by calculating a correlation matrix including all of the original coordinates of the instances in the dataset, and finding its eigenvectors on diagonalising this matrix. Thus, the eigenvectors are the axes in this transformed space, and can be ranked according to their eigenvalues, which reflect the variance across each axis. By selecting the top  $n$  eigenvectors, the multi-dimensional space is reduced to  $n$  (typically two-dimensional, to aid visual representations on paper). Given how the top  $n$  eigenvectors are selected by the method, these explain the maximum variability possible with only  $n$  vectors.

However, in the MOOC data, there is an assumption the method relies on which might be violated: that there should not be significant *outliers*, as they would have a disproportionate influence on the results. MOOC participation typically follows a long tail (with the majority of the participants showing low engagement, and very few showing extremely high engagement). My explorations with this method using WEKA’s implementation confirmed this. Appendix I shows the results using all features in two



runs' datasets (specifically run one of Understanding Language and run six of Portus, the largest and the smallest of the MOOCs, respectively).

Whilst still useful for identifying the most relevant features for each dataset (as listed in page I-1), constructing a two-dimensional space from the top two eigenvectors is to be discouraged in this context. The resulting attributes explain only a very small proportion of the variability observed in each case (e.g. 27% in run one of Understanding Language, and 32.1% in run six of Portus<sup>18</sup>). Even increasing the number of dimensions would not improve matters significantly, as incorporating the top five ranked attributes would only explain 39% and 47.6% of the whole variability in these two respective courses. These are not acceptable values as it would be desirable to set a number of dimensions that can explain around 90% of the variability (or over 60% at the very least, as ), otherwise no valuable insights would be obtained from the subsequent clustering.

Further, the fact that the eigenvectors chosen by PCA are potentially different for each run of these MOOC is an important limitation for inter-run comparisons as well as comparisons across MOOCs, let alone comparing against other peer-supported environments as per the aim of this thesis. Finally, the resulting clusters (should we use the "ranked attributes" that PCA outputs as the reduced feature set as input to the clustering) would have very low-interpretability as the features themselves are a linear combination of a subset of the original feature set, and therefore, unsuitable for this application, where high-interpretability is desired in clusters.

Given these considerations, I did not pursue this line of enquiry in what follows, despite being one commonly used for dimensionality reduction as explained above.

## 5.5.2 Semantically-chosen features

An alternative approach for dimensionality reduction of the feature space is to group features semantically to investigate how expressive they are for the data. Choosing a feature set that expresses well the differences amongst instances in the data would support the generation of clusters that are a good fit for the data, as well as being of high interpretability. Therefore, I compared the subsets of features described in Section 5.4: dialogic features, interval features and badge features, and performed experiments with the commonly used *k-Means* clusterer (applied by [Kizilcec et al. \(2013\)](#); [Ferguson and](#)

---

<sup>18</sup>These percentages are calculated from the highlighted values in Appendix I, after subtracting them from 1 to get the cumulative variance explained with the inclusion of a given factor and all higher-ranked factors. Including the top two ranked attributes as calculated by PCA explains  $(1 - 0.679) * 100\%$  of the variability in the six run of Portus. Hence, an explained variability of only 32.1%.

Clow (2015a,b); Tseng et al. (2016); Dowell et al. (2018); Kizilcec and Chen (2020)) applied to each of these subsets.

Figure 5.4 shows the performance of the k-Means clusterer using the dialogic features from Section 5.4.1, with the number of clusters ( $k$ ) varying between 2 and 10. It can be seen that the within-cluster sum of squared errors decreases as  $k$  increases. This is an indication that the instances in each cluster are more similar to each other as the variation tends to zero. In the limit, where all the instances in each cluster are identical (or there is exactly one cluster per instance), the differences between each and the means (characterised by the coordinates of the centroid) are exactly zero and therefore the sum is zero. Particularly in Figure 5.4, it can be seen that the decrease slows down after  $k = 3$  (the ‘elbow’ for the clustering), suggesting that this would be the optimal number of clusters for both Understanding Language and Portus.

The performance of the k-Means clusterer is much poorer with interval features (from Section 5.4.2), judging by the within-cluster sum of square errors which is in its thousands as seen in Figure 5.5. In terms of the ‘elbow’, there seem to be two potential candidates, in  $k = 3$  and  $k = 8$ . However, given that the algorithm performs much better with dialogic features, it would seem that interval features are not as expressive to differentiate instances. Similarly, when performing the same set of experiments on badge features from Section 5.4.3, within-cluster sum of square errors starts very high, with values in the tens of thousands, and then drops to zero at exactly  $k = 3$ . This can be seen in Figure 5.6. Whilst a small within-cluster sum of square errors is desirable, it plummeting to zero this early means that all of the instances in the cluster have exactly the same values as the cluster centroids, and therefore the feature set is not sufficiently expressive for this data. Therefore, it would be very unlikely that any nuances of behaviour would be noticed by the clusterer.

Given these observations, I focus on the use of dialogic features in what remains of this thesis.

## 5.6 Clustering algorithm on MOOC features

Having chosen to do the clustering on dialogic features only, the next step is to confirm whether the widely used *k-Means* clustering algorithm is still a good choice for this dataset, and whether  $k = 3$  (found by the Elbow method as seen in Figure 5.4) is a good choice for number of clusters.

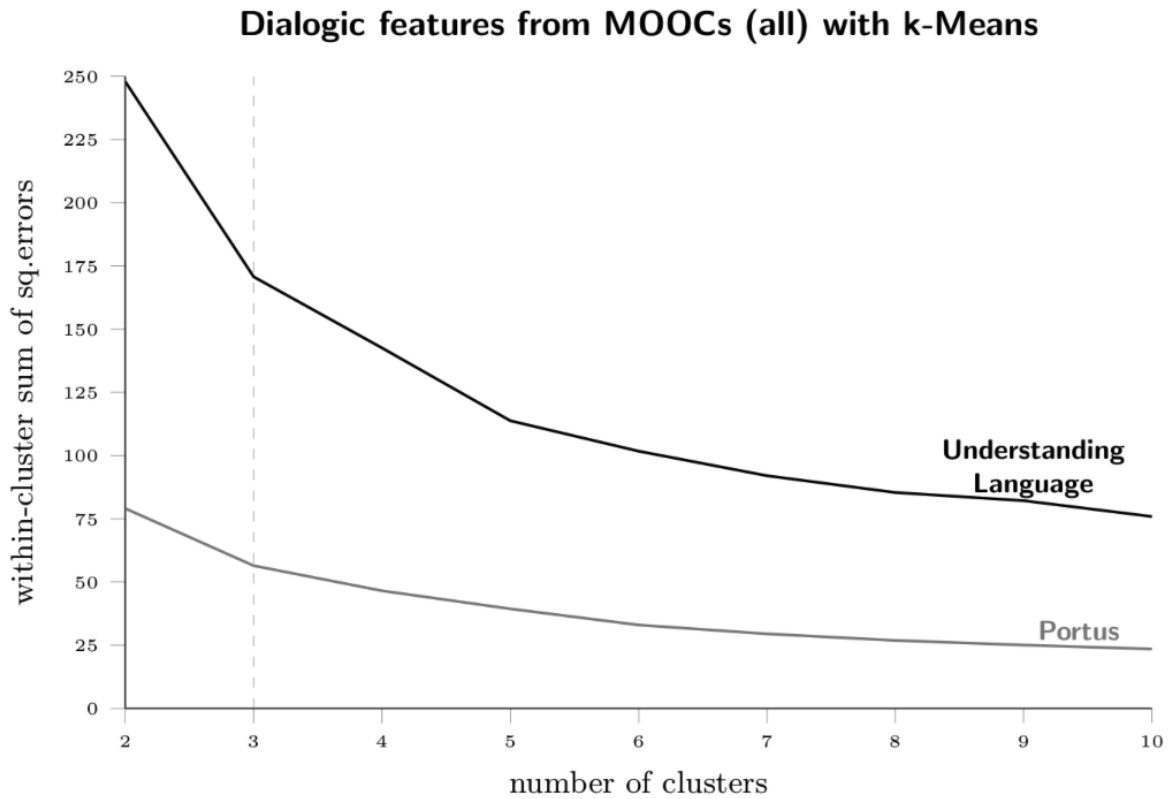


FIGURE 5.4: Inspecting the within-cluster sum of square errors to assess the performance of k-Means using only dialogic features on data from learners in for Understanding Language (all runs) and Portus (all runs).

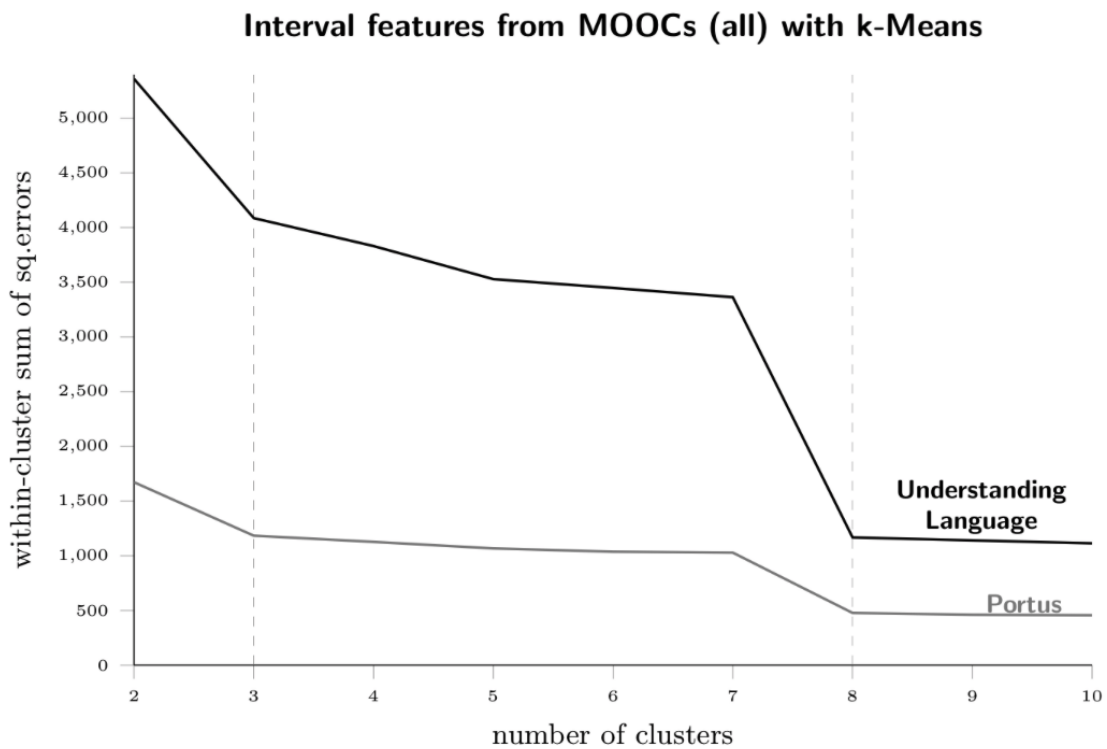


FIGURE 5.5: Inspecting the within-cluster sum of square errors to assess the performance of k-Means using only interval features on data from learners in for Understanding Language (all runs) and Portus (all runs).

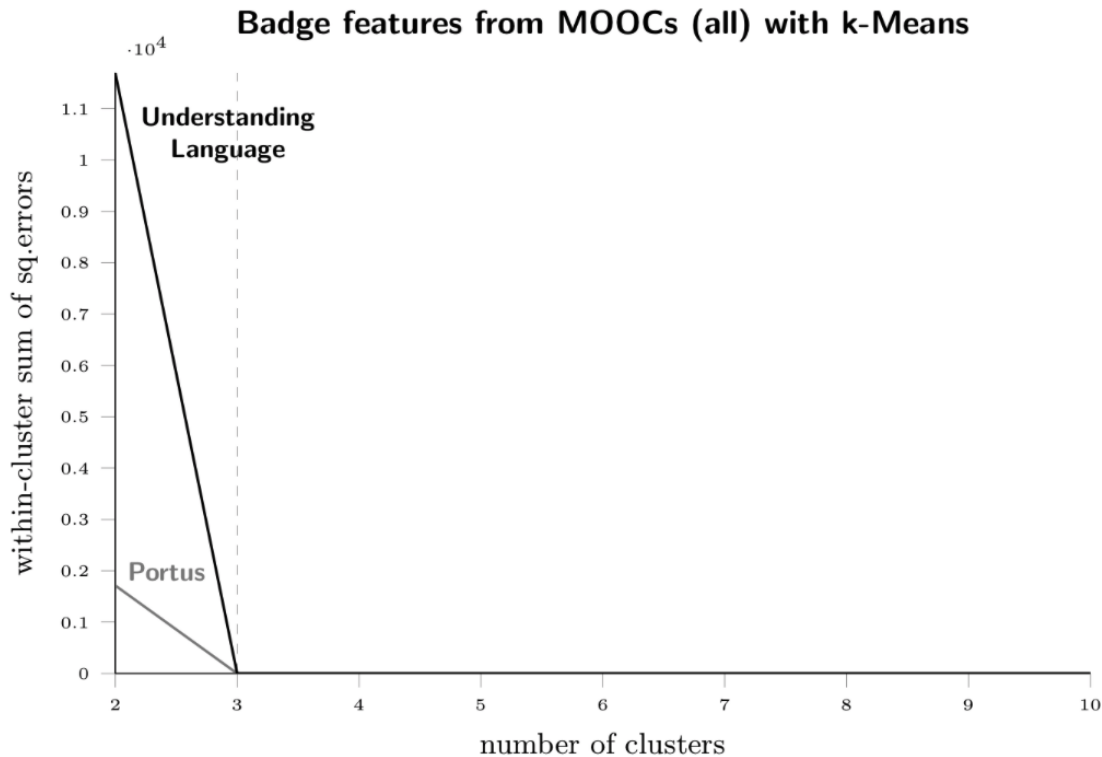


FIGURE 5.6: Inspecting the within-cluster sum of square errors to assess the performance of k-Means using only badge features on data from learners in Understanding Language (all runs) and Portus (all runs).

### 5.6.1 Which clustering algorithm?

To select an appropriate clustering algorithm, in addition to the k-means clustering algorithm, which was used to select the reduced feature set, I consider two alternatives and compare their performance. The first clusterer is Expectation Maximisation (EM), also mentioned in Table 2.3 as having been used by [Bogarín et al. \(2014\)](#) and others; and the second one is X-Means ([Pelleg and Moore, 2000](#)), given that it is considered a good improvement to k-Means. Like in the experiments just presented in Section 5.5, all runs in each of both MOOCs were collated into two large instance files upon which the clustering algorithms were applied.

However, in terms of assessing the performance, the within-clusters sum of squared errors could no longer be used as a metric for comparison, as this is not used by EM or X-Means<sup>19</sup> An approach that can be useful in such cases is to preprocess the data with a given clustering algorithm as a filter, such that to the existing feature set, a new feature named “cluster” is added to each instance, indicating the label to the cluster it

<sup>19</sup>The ‘log-likelihood’ is an output for EM and X-Means but not for k-Means, so it cannot be used as a comparative performance metric either.

was assigned to by the algorithm<sup>20</sup>. The resulting file is subsequently subjected to a supervised learning algorithm with cross-validation (a classifier, such as the scikit-learn `DecisionTreeClassifier`), using the newly-created feature “cluster” as the ground truth, or predictive class. The percentage of incorrectly clustered instances that the classifier outputs is therefore a suitable measure of the goodness of the clusterer, as the classifier only sees the feature values to predict the labelled cluster, and if the within-cluster similarity is high, the classification will be highly accurate. Figures 5.7 and 5.8 show that X-Means outperforms both k-Means and Expectation Maximisation (EM) in both Understanding Language and Portus MOOCs datasets with dialogic features.

### 5.6.2 How many clusters?

In order to perform the comparison, once again I varied the number of clusters  $k$  which is a parameter for k-Means. Neither EM nor X-means require this parameter, as they are able to find the number of clusters for optimum fitness from the data if not provided. However, in both cases, it is possible to set upper and lower bounds for the number of clusters, and when setting both values to a given  $k$ , the algorithms are forced to group the instances into that given number of clusters.

Interestingly, when left alone (i.e. when  $k$  is not set), X-Means typically returns seven clusters on this data. When running the classifier with a ten-fold validation testing upon the seven clusters given by X-Means the detailed accuracy by class is very high for both Understanding Language (all runs) and Portus (all runs too). The information retrieval values are very high across all classes as shown in Appendix J. However, when running the clusterer on separate runs of the MOOCs, on occasion it would return four clusters instead (particularly when the clusterer was used as an instance filter). The reason behind it might be that the difference in performance is almost negligible across the different values of  $k$ , as shown in both Figures.

Therefore, and for the sake of consistency, I chose to use seven clusters in the subsequent analysis. Seven clusters are also closer to Chua’s taxonomy in terms of number of distinctively identifiable groups. In addition, in the particular context of dialogic engagement, two distinct categories contain the majority of the learners of the datasets (i.e. the *asocial learners*, who did not post any comments, and the *loners*, who did not engage in conversations despite having posted comments). Grouping everyone else under the same category, in a third, final cluster, did not seem of sufficient interest as it would not provide insights on the nuanced behaviours occurring in the long tail of the

---

<sup>20</sup>In WEKA the cluster names assigned are: cluster1, cluster2, ..., cluster $k$ .

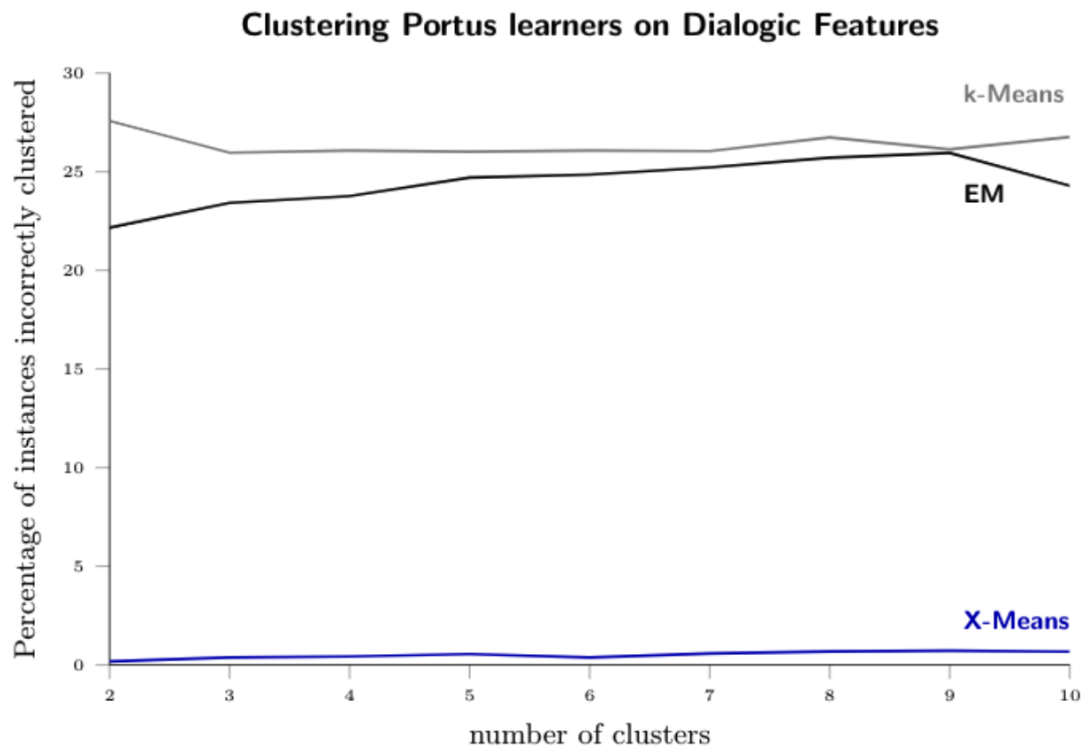


FIGURE 5.7: Percentage of instances incorrectly clustered according to the scikit-learn `DecisionTreeClassifier` with varying values of  $k$ , using only dialogic features on data from learners in Portus (all runs).

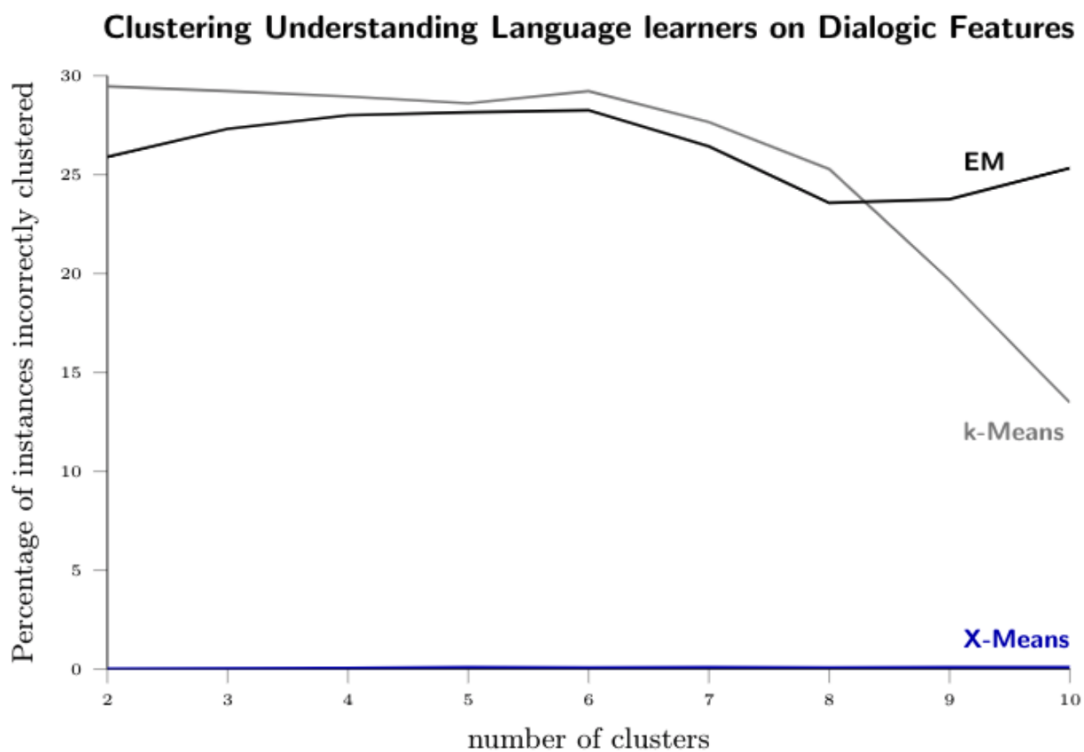


FIGURE 5.8: Percentage of instances incorrectly clustered according to the scikit-learn `DecisionTreeClassifier` with varying values of  $k$ , using only dialogic features on data from learners in Understanding Language (all runs).

datasets. Therefore, a more interesting clustering for the data used here would necessarily need to be of a larger number of clusters.

However, it is important to remember that many researchers report findings with a very small number of clusters. For example, [Tseng et al. \(2016\)](#) report having found three distinct classes of learners (active, passive, and bystanders); [Kizilcec and Chen \(2020\)](#), found three clusters (of low, medium and high levels of engagement activity); and [Bogarín et al. \(2014\)](#) also identify three clusters to characterise failing students and two types of passing students. Several others in [Table 2.3](#) do report categorisations of learners with very few classes. Therefore this is also a valid choice, in particular because the clusterer offered classes with very good precision and recall even with only four clusters, as per the results shown in [Appendix K](#). This question will be revisited in [Section 5.7](#), in particular, when discussing [Table 5.16](#).

## 5.7 Results

Once decisions were made regarding the selection of a reduced set of engineered features (dialogic features), a clustering algorithm (X-Means) and a suitable number of clusters ( $k = 7$ ), the next step is to interpret the results from the clustering process for these selections. In particular, in this section I investigate the size of the resulting clusters and their coherence (in [Section 5.7.1](#)); give them meaningful names based on the central measures of the instances within for both MOOCs when all the runs are aggregated (in [Section 5.7.2](#)), and repeat the process for each run of each MOOC (in [Section 5.7.3](#)). Then, in [Section 5.7.4](#), I look into the distribution of learners across these newly-named categories and compare these distributions against those given through the heuristic by [Chua et al. \(2017\)](#), as presented in [Section 5.3.1](#).

### 5.7.1 Size and coherence of resulting clusters

I investigate the size of the resulting clusters and their coherence. In order to do so, it is necessary to inspect the confusion matrices generated as described in [Section 5.6.1](#), when the scikit-learn `DecisionTreeClassifier` was used to determine the goodness of the fit. [Figures 5.9](#) and [5.10](#) show the confusion matrices for the classifier when predicting the seven clusters found by X-Means on the datasets containing all of the runs available for the Portus MOOC and the Understanding Language MOOC, respectively.

A confusion matrix plots the goodness of a classifier by showing whether the ground truth (commonly depicted in the  $x$  axis) matches the predicted class (in the  $y$  axis).

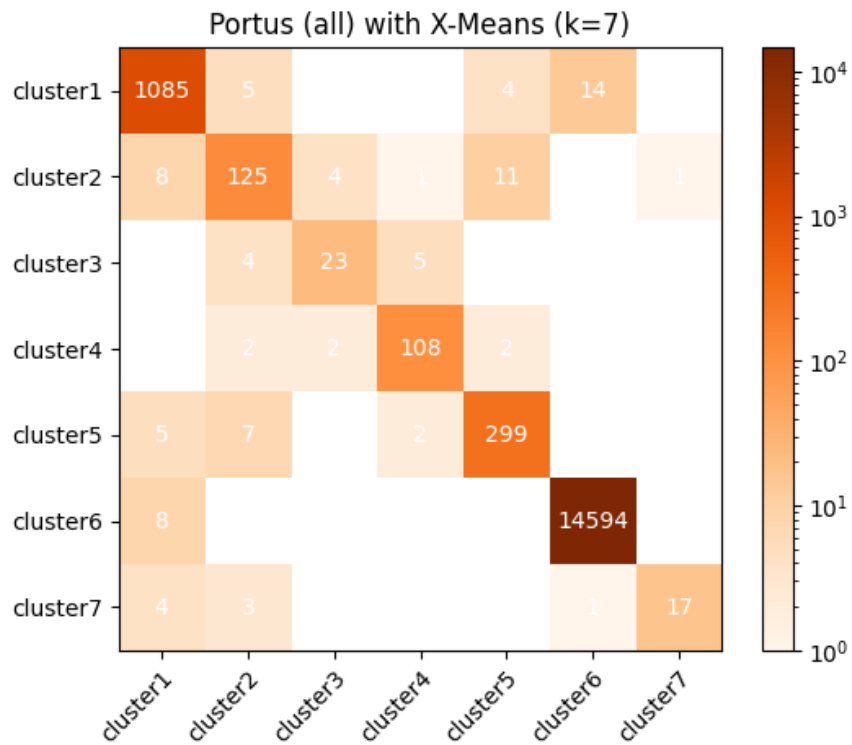


FIGURE 5.9: Confusion matrix plots for the scikit-learn DecisionTreeClassifier on the seven clusters found by X-Means applied to the Portus MOOC (all runs combined), with  $k = 7$

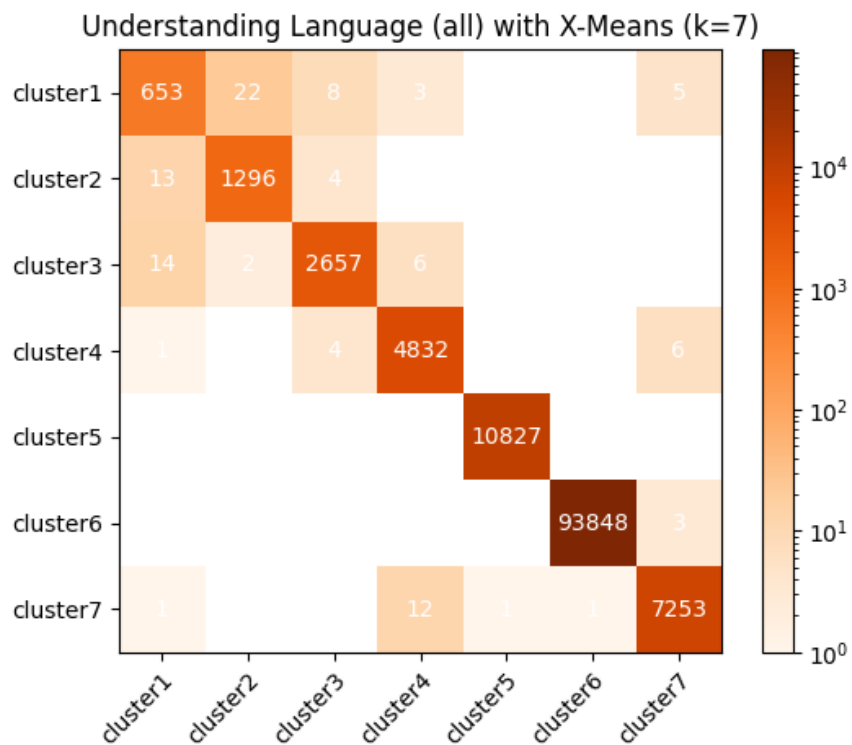


FIGURE 5.10: Confusion matrix plots for the the scikit-learn DecisionTreeClassifier on the seven clusters found by X-Means applied to the Understanding Language MOOC (all runs combined), with  $k = 7$



The numbers of instances for which the prediction is correct lie on the diagonal of the matrix. Misclassifications, lying elsewhere on the matrix when occur, are also important to inspect to gain insight on the kinds of “confusion” that may take place between classes. This is evident in the matrices shown in Figures 5.9 and 5.10, where, in addition, I used a heatmap to highlight the intensity of the matches (the darker the colour, the higher the number of instances). Specifically, I used a logarithmic-scale colour map to facilitate its reading, due to the high imbalance of the clustered data, as indicated by the colour bar next to each confusion matrix.

On closer inspection, it is possible to make the following observations: firstly, as mentioned, the clusters found are very imbalanced. In particular, cluster 6 in both cases, contain the highest number of instances by far, both with the largest numbers of correctly classified instances (14,594 and 93,848 respectively), several orders of magnitude more than the other classes, which are nonetheless largely accurately classified (though there is some noise outside the diagonal indicating a few dozen misclassifications). The great class imbalance observed is expected, given the nature of the data, and the 90-9-1 rule as discussed in Section 5.3.

Despite this large imbalance, the clusters are inherently coherent, as the high level of similarity between instances within the same cluster are usually correctly predicted when subjected to the scikit learn `DecisionTreeClassifier`. This is evident in Figures 5.9 and 5.10, since both matrices are markedly diagonally-dominant, indicating a high coherence in the classes, which is most particularly true for the largest classes in both sets of results (containing 14,594 and 93,848 instances). Given this high inter-cluster coherence, the next logical step is to assign clusters meaningful names that are able to characterise each cluster in domain-interpretable terms.

## 5.7.2 Semantically chosen names for clusters in both MOOCs

Having established that the seven clusters found via unsupervised learning are coherent, I next inspect the resulting clusters to give them meaningful names based on the central measures of the instances within.

Box-and-whiskers plots, such as those in Figures 5.11 and 5.12, are used to demonstrate central measures and dispersion of the data and are interpreted as follows: the green triangle indicates the mean, the orange bar indicates the median. The lower and upper limits of the boxes represent the 25<sup>th</sup> and 75<sup>th</sup> centiles, otherwise known as the interquartile range (IQR). The ‘whiskers’, are the lines extending to either side of the box to 1.5 times the IQR, to indicate the variation in the data. Any other values outside

this range are deemed as outliers and are indicated by circles. The number of comments are plotted against a logarithmic scale.

### Portus (all runs)

For each cluster in the Portus dataset containing all runs of the MOOC, their semantics are loosely based on the names of the groups by the heuristic in [Chua et al. \(2017\)](#), according to the mean and median values for the dialogic features as shown in Figure 5.11. These are as follows:

**cluster1** *Initiators without replying:* This group has non-zero starting posts (SP), with both mean and median greater than one; non-zero lone posts (LP), with a mean and median of ten lone posts and non-zero first replies (FR). However, both the mean and median are zero for initiators replies (IR), and additional replies (AR). There are 1,108 instances in this cluster.

**cluster2** *Active social learners:* In this group, the central measures for both SP and LP are greater than 10, and all FR, IR, AR are greater than one. There are 150 instances in this cluster.

**cluster3** *More active social learners:* As above but even higher number of comments for the means and medians of all dialogic features. There are 32 instances in this cluster.

**cluster4** *More active social learners who do not give additional replies:* Similar to cluster5 below but with an even higher level of comment activity. All dialogic features have central measures greater than one, except AR, for which is zero in both counts. This group has 114 instances.

**cluster5** *Active social learners who do not give additional replies:* All dialogic features have central measures greater than zero, except AR, for which is zero in both counts. There are 313 instances in this cluster.

**cluster6** *Asocial learners:* All dialogic features are at zero, except for the outliers that are shown. There are 14,602 instances in this cluster.

**cluster7** *Initiators who respond.* This group has non-zero starting posts (SP) yet a zero median for initiators replies (IR). Though the mean is non-zero, 75% of the members of this cluster have less than one IR. Given the size of the cluster, the mean is very sensitive to the outliers' effect (there are 25 instances in this cluster, and one of them has given tens of initiator's replies.)

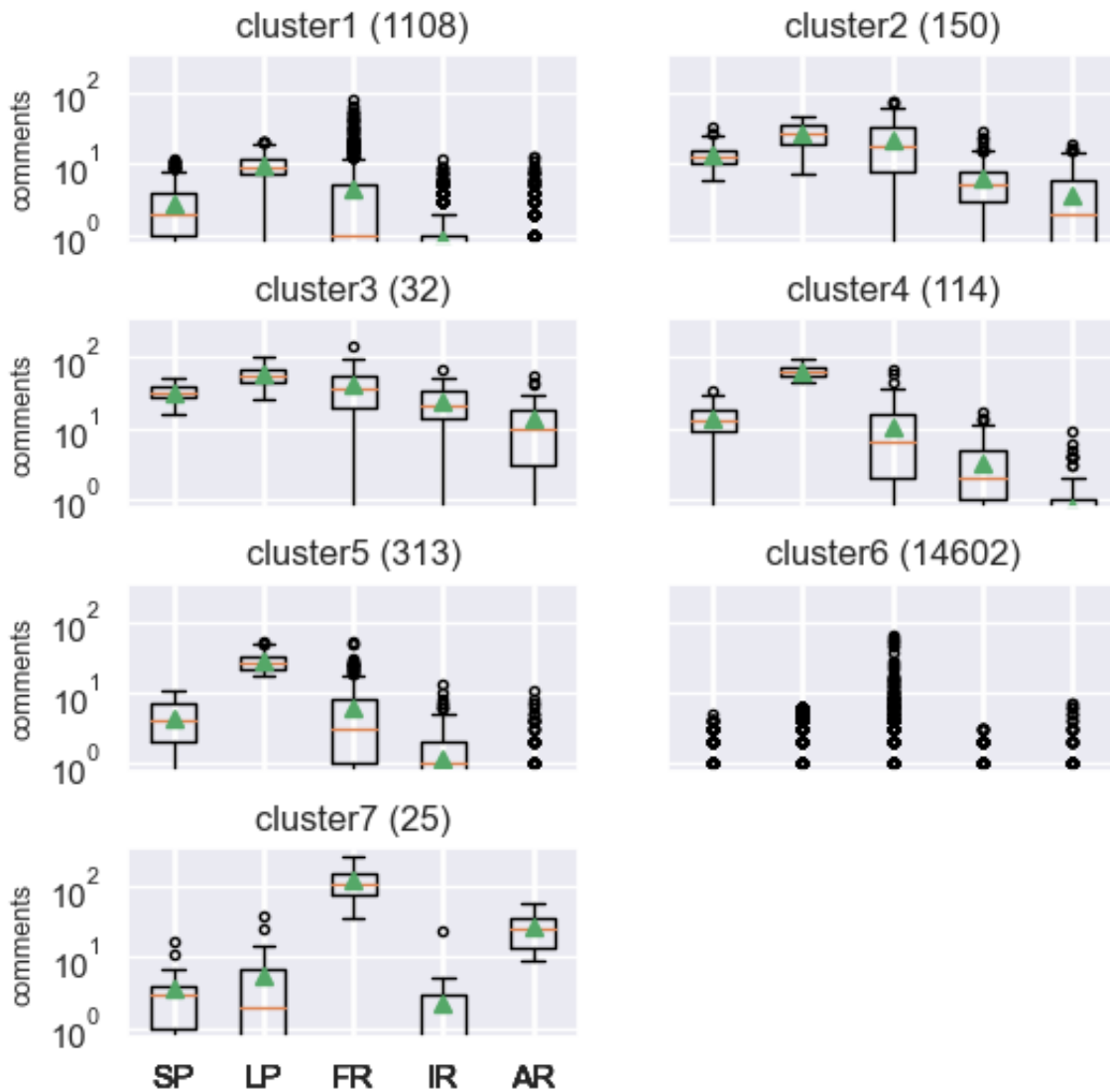


FIGURE 5.11: Box-and-whisker plots for the dialogic features on clusters found by the X-Means clustering algorithm on the Portus MOOC (all runs), with  $k = 7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1**: Initiators without replying; **cluster2**: Active social learners; **cluster3**: More active social learners; **cluster4**: More active social learners who do not give additional replies; **cluster5**: Active social learners who do not give additional replies; **cluster6**: Asocial learners; **cluster7**: Initiators who respond.

## Understanding Language (all runs)

For Understanding Language, I analysed the central measures of the clusters shown in Figure 5.12, guided by the categories in Chua et al. (2017) as explained in Section 5.3 and similarly applied above. This resulted in the following cluster names:

- cluster1** *Active social learners*: In this group, the central measures for both SP and LP are greater than ten, and all FR, IR, AR are greater than one. There are 691 instances in this cluster.
- cluster2** *More active social learners who do not give additional replies*: This is similar to cluster3 below but with a higher level of comment activity. All dialogic features have central measures greater than zero, except AR, which has a median of zero and a mean less than one. All learners in this category have zero AR, except outliers. There are 1,313 instances in this cluster.
- cluster3** *Active social learners who do not give additional replies*: Similar to cluster2, only with a lower level of comment activity. There are 2,679 instances in this cluster.
- cluster4** *Active social learners without turn-taking*: Non-zero SP, LP and FR, but zero for IR and AR for most instances except outliers (no initiators replies, and no additional replies). There are 4,843 instances in this cluster.
- cluster5** *Loners*: All features are zero apart from LP for most of the 10,827 instances (except for outliers).
- cluster6** *Asocial learners*: All dialogic features are zero. In the plot, only outliers are shown. The size of this cluster is 93,851 instances.
- cluster7** *Initiators without replying*: This group has non-zero starting posts (SP) yet a zero median for initiators replies (IR). Though the mean is non-zero, 75% of the members of this cluster have less than one IR. There are 7,268 instances in this cluster.

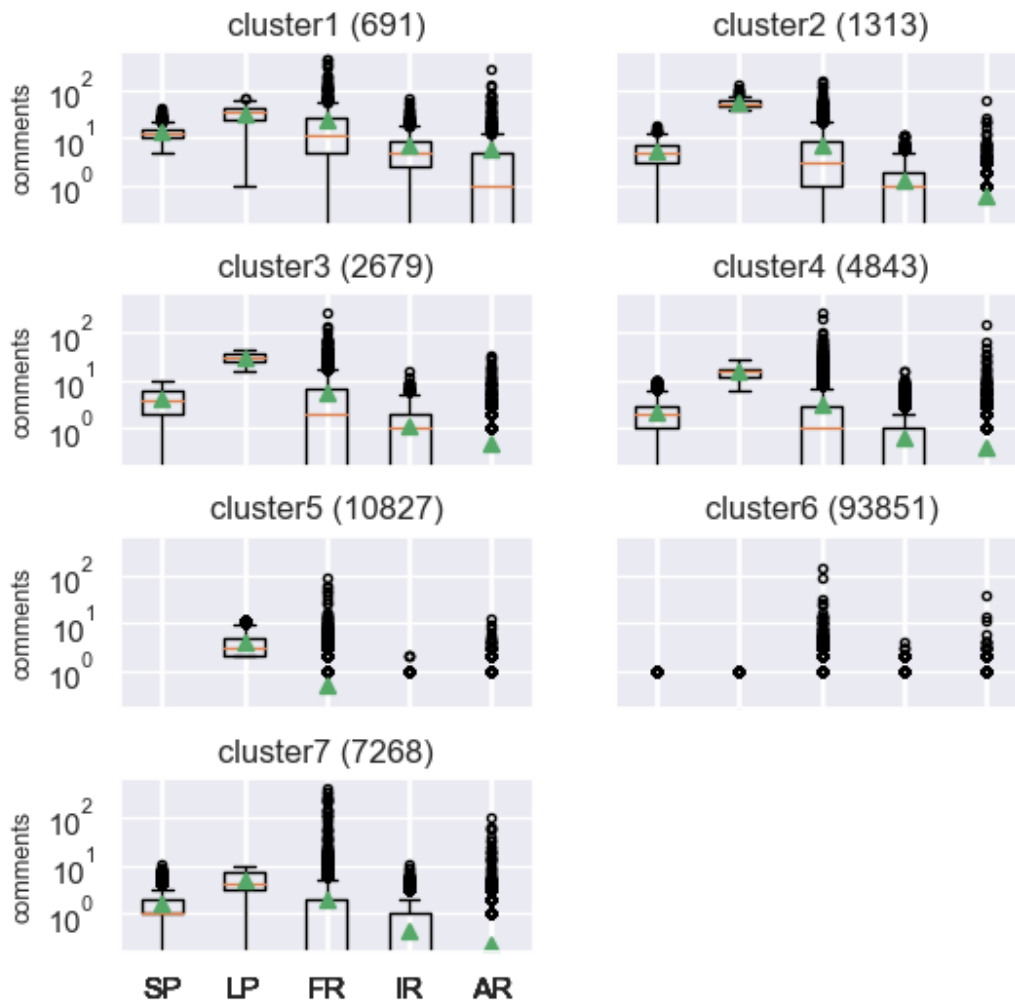


FIGURE 5.12: Box-and-whisker plots for the dialogic features on clusters found by the X-Means clustering algorithm on the Understanding Language MOOC (comprising all of the available runs), with  $k = 7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1** : Active social learners; **cluster2** : More active social learners who do not give additional replies; **cluster3** : Active social learners who do not give additional replies; **cluster4** : Active social learners without turn-taking; **cluster5** : Loners; **cluster6** : Asocial learners; **cluster7** : Initiators without replying.

### 5.7.3 Semantically chosen names for clusters in each run of a MOOC

As mentioned in Section 5.2.1, even small variations in the learning design between consecutive runs may lead to variations in the evidenced learning engagement. This is investigated by repeating the process I have just described in Section 5.7.2, but instead inspect the confusion matrices and the box-and-whiskers plots associated to *each* of the runs of each MOOC. All of these plots are available in Appendix L, comprising sixteen confusion matrices and sixteen groups of seven box-and-whiskers plots.

It is important to note that, using those plots, I applied the same process for determining the semantics of each cluster as in the previous section, that is, naming the clusters as closely as possible<sup>21</sup> to the classes provided by Chua et al. (2017), whilst incorporating any relevant variation (such as a higher level of activity, or an additional distinction that would count as a sub-category of those original classes).

However, to ease the identification of the classes found across all of the sixteen runs' studies (six for Portus and ten for Understanding Language), I added another modifier to the name, which is pre-pending a number followed by a letter, to facilitate keeping track of subdivisions of the original categories made by Chua et al. (2017), especially given there is a high similarity amongst several of these names. These are shown under the column "Classes found" in the set of Tables 5.12 to 5.15. In particular, Table 5.12 shows the mapping between the cluster name as assigned by the X-Means clusterer and the associated category name found through manual inspection of the boxplots (for all of the runs of Portus), whereas Table 5.13 shows the size of each of those clusters in number of instances. The remaining two tables show the same findings but for Understanding Language.

---

<sup>21</sup>Though I named some of the categories similarly to those in Chua et al. (2017) these are not exactly the same. For example, an *initiator without replying* "never replied others' initiating posts despite receiving replies from others" (Chua et al., 2017). This is too tight a definition, given the overall behaviour as observed in the cluster containing 1,098 learners in Figure K.1, where with the exception of a handful of outliers, for all cluster instances AR=0, and both IR and FR had very low values too. This is even more evident in Figure K.2, where the median of FR considering all 9,279 learners in this cluster is exactly one, and for both IR and AR is exactly zero.

TABLE 5.12: Semantic classes for the clusters found by X-Means in Portus (for each run), with  $k = 7$ .

		Portus					
Classes found		1	2	3	4	5	6
1-	asocial learners	cluster6	cluster7	cluster7	cluster7	cluster7	cluster5
2-	loners			cluster5	cluster1		cluster4
2a-	more active loners						
3-	initiators without replying	cluster1	cluster5	cluster4			
3a-	more active initiators without replying						
4-	initiators who respond	cluster7					
5-	replier		cluster2	cluster3		cluster6	
6-	reluctant ASL	cluster3				cluster5	
7-	ASL without turn-taking	cluster5	cluster6	cluster2	cluster5		cluster6
7a-	more active SL without turn-taking						
7aa-	even more active SL without turn-taking						
8-	active social learners	cluster2	cluster4	cluster1	cluster3	cluster4	cluster7
8a-	more active social learners	cluster4	cluster1	cluster6	cluster4	cluster3	cluster3
8aa-	even more active social learners				cluster2		
8b-	ASL who do not give additional replies		cluster3		cluster6	cluster2	cluster2
8bb-	more active SL who do not give additional replies					cluster1	cluster1

TABLE 5.13: Numbers of learners in each of the semantic classes for the clusters found by X-Means in Portus (for each run), with  $k = 7$ .

		Portus					
Classes found		1	2	3	4	5	6
1-	asocial learners	4134	3562	1418	2107	2000	848
2-	loners			60	177		45
2a-	more active loners						
3-	initiators without replying	174	286	31			
3a-	more active initiators without replying						
4-	initiators who respond	28					
5-	replier		5	6		2	
6-	reluctant ASL	589				115	
7-	ASL without turn-taking	53	81	11	47		30
7a-	more active SL without turn-taking						
7aa-	even more active SL without turn-taking						
8-	active social learners	589	53	25	75	44	8
8a-	more active social learners	11	12	5	7	7	2
8aa-	even more active social learners				17		
8b-	ASL who do not give additional replies		34		26	62	22
8bb-	more active SL who do not give additional replies					21	14

TABLE 5.14: Semantic classes for the clusters found by X-Means in and Understanding Language (for each run), with  $k = 7$ .

Classes found		Understanding Language									
		1	2	4	5	6	7	8	9	10	11
1-	asocial learners	cluster7	cluster7	cluster6	cluster7	cluster6	cluster6	cluster5	cluster7	cluster7	cluster7
2-	loners	cluster4	cluster6	cluster5	cluster4	cluster1	cluster7	cluster4	cluster5	cluster5	cluster5
2a-	more active loners		cluster3								cluster2
3-	initiators without replying	cluster2	cluster2	cluster2			cluster3	cluster2			
3a-	more active initiators without replying	cluster1		cluster7							
4-	initiators who respond	cluster3	cluster1								
5-	replier										
6-	reluctant ASL										cluster6
7-	ASL without turn-taking	cluster6	cluster4	cluster1	cluster2	cluster4	cluster5	cluster1	cluster3	cluster6	cluster1
7a-	more active SL without turn-taking			cluster4	cluster3	cluster7	cluster1	cluster7	cluster2	cluster4	cluster3
7aa-	even more active SL without turn-taking										cluster1
8-	active social learners	cluster5	cluster5	cluster3	cluster5	cluster2	cluster4	cluster6	cluster4	cluster3	cluster4
8a-	more active social learners					cluster5	cluster2	cluster3	cluster6		
8aa-	even more active social learners										
8b-	ASL who do not give additional replies				cluster6	cluster3			cluster1	cluster2	
8bb-	more active SL who do not give additional replies				cluster1						

TABLE 5.15: Numbers of learners in each of the semantic classes for the clusters found by X-Means in Understanding Language (for each run), with  $k = 7$ .

Classes found		Understanding Language									
		1	2	4	5	6	7	8	9	10	11
1-	asocial learners	22108	15924	9457	9782	4832	7315	2905	5097	2613	3432
2-	loners	2936	2385	679	661	312	651	177	399	264	378
2a-	more active loners		848								142
3-	initiators without replying	920	436	876			204	62			
3a-	more active initiators without replying	899		248							
4-	initiators who respond	431	193								
5-	replier										
6-	reluctant ASL										5
7-	ASL without turn-taking	503	497	226	125	56	100	37	149	89	75
7a-	more active SL without turn-taking			136	108	53	64	24	81	46	49
7aa-	even more active SL without turn-taking										22
8-	active social learners	161	153	95	6	28	87	9	17	20	21
8a-	more active social learners					2	27	2	3		
8aa-	even more active social learners										
8b-	ASL who do not give additional replies				211	64			50	14	
8bb-	more active SL who do not give additional replies				55						



## 5.7.4 Distribution of learners across the newly-named clusters

This section analyses the distribution of learners across the classes identified in the previous section. I perform this analysis in a similar manner as that presented in Section 5.3.1, that is, by inspecting bar charts plotting the numbers of learners on each category.

However, in order to illustrate the importance of having renamed<sup>22</sup> the clusters found by X-Means through a manual inspection of their semantics, let us first consider the numbers of learners assigned cluster labels given by the algorithm, ignoring for a moments the names I gave them by inspecting the boxplots in Appendix L, as explained.

It is clear that these would not easily interpretable, other than signalling that there is a large class imbalance observed across all runs, in which some clusters have several thousands of learners whereas for others it is merely a handful. This large class imbalance is evidenced in both MOOCs. For example in Portus (in Table 5.13, cluster6 in run one (shown in orange) and cluster7 in run two (in red) have around four thousand learners each, whereas cluster3 in run six (in cyan) and cluster6 in run five (in orange still) have only two learners each. This is also noticeable in Understanding Language (in Table 5.15, where for cluster7 in run one (shown in red) there are 22,108 learners, yet only two in cluster5 for run six (in yellow).

The difficulty in gaining further insights from this clustering lies with the fact that the X-Means algorithm groups instances according to the similarity found in the defining features in each experiment (i.e. for each separate run of the MOOCs). Though the clusters are highly coherent (as shown in the confusion matrices in Figures L.1 and L.8), there is no shared memory across experiments, and therefore the cluster labels are freshly assigned each time and cannot be expected to be coherent across experiments as they are within.

The manual inspection of the box-and-whisker plots for each cluster is one way to overcome this evident difficulty of comparing clusters across different runs of the studied MOOCs. By relabelling the clusters found and grouping them by their semantic names before making the bar charts (as described at the start of this section), there will be an improved interpretability of the variations in groups composition amongst runs of each MOOC. This is explained in the next section.

---

<sup>22</sup>Another note on names. In earlier drafts of this thesis I had used different names for several of these categories but, for the sake of clarity in the comparisons, I reverted to the original names given in Chua et al. (2017). In particular, and as mentioned before, I did not want to use the term ‘loner’.

### Replotting bar charts to include every semantic class

The bar charts that include every semantic class listed in Tables 5.13 and 5.15 are shown in Figures 5.13 and 5.14. There are a number of insights emerging from these charts. Firstly, though many different classes are found within the MOOCs studied, only a few appear consistently in most runs (if not all). For Portus (Figure 5.13), these are:

- ‘1-asocial learners’,
- ‘7-ASL without turn-taking’,
- ‘8-active social learners’,
- ‘8a-more active social learners’.

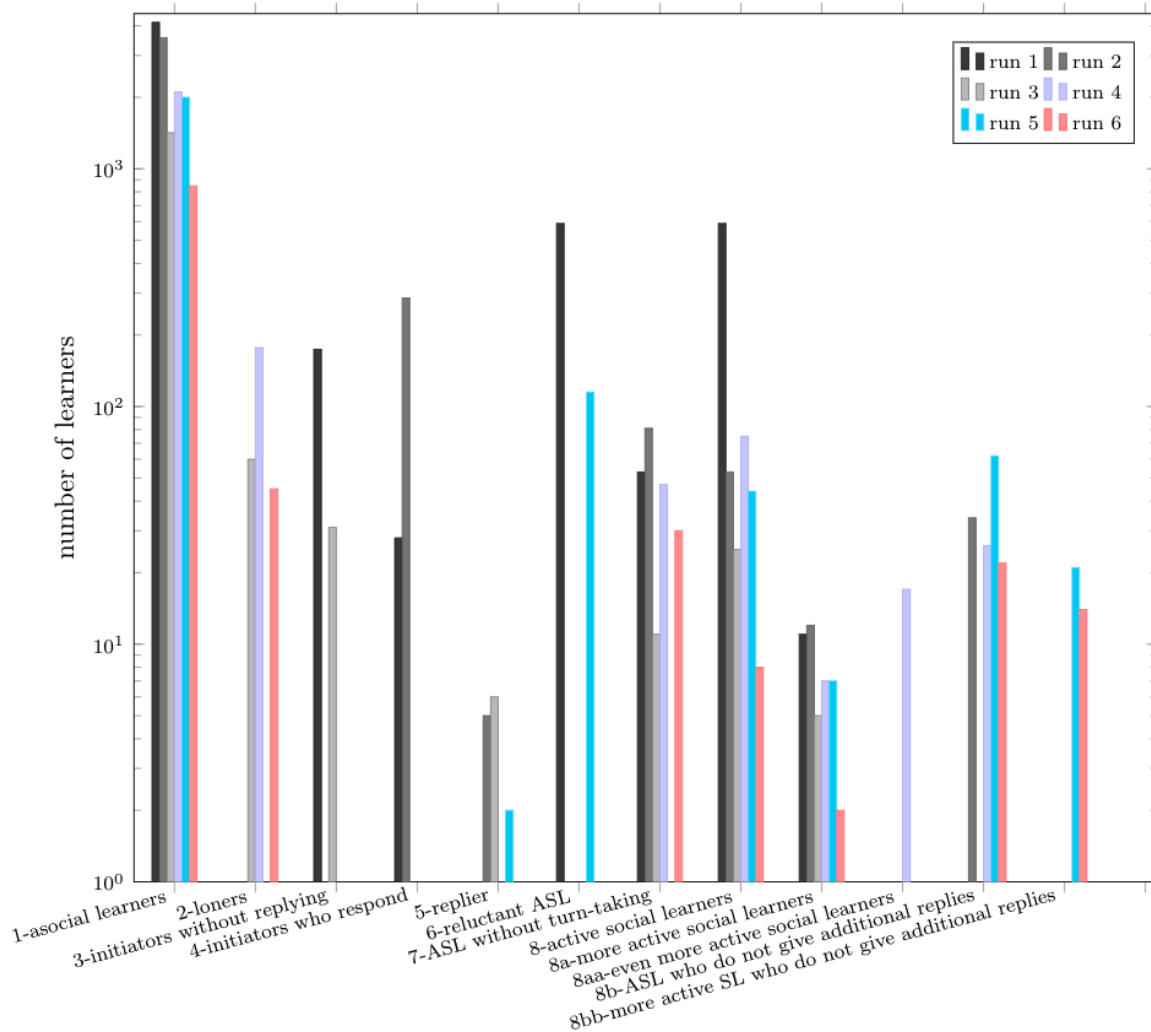


FIGURE 5.13: Distribution of social learners across on Portus according to the clusters found by X-Means ( $k = 7$ ) and interpreted using the classification by Chua et al. (2017).

whereas for Understanding Language (Figure 5.14), these are:

- ‘1-asocial learners’,
- ‘2-loners’,
- ‘7-ASL without turn-taking’,
- ‘8-active social learners’.

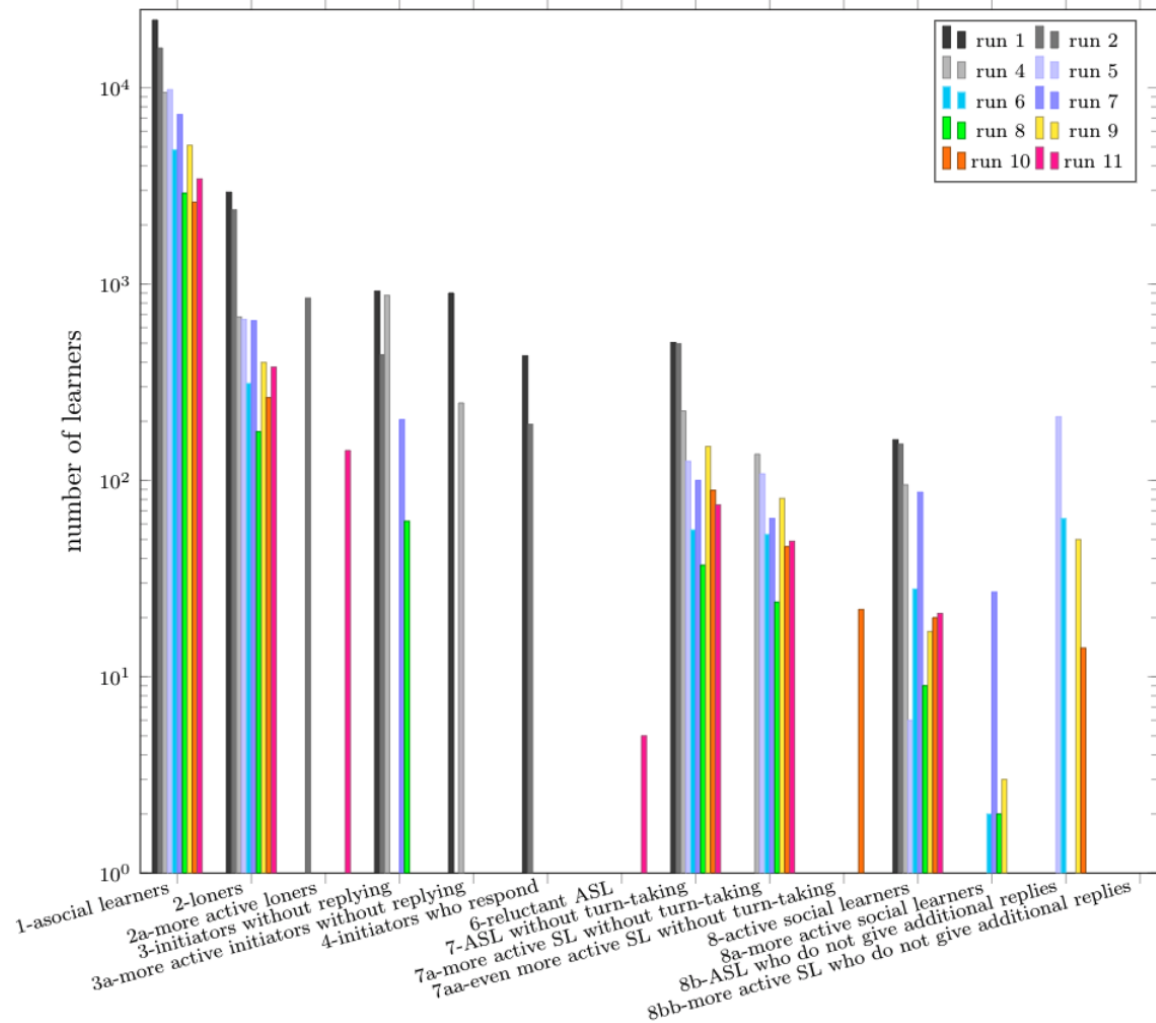


FIGURE 5.14: Distribution of social learners across on Understanding Language according to the clusters found by X-Means ( $k = 7$ ) and interpreted using the classification by [Chua et al. \(2017\)](#).

Also noteworthy are variations in the level of a given type of activity are often found, these are not observed consistently across all runs. For example, run two in Understanding Language had a ‘2a-more active loners’ class, which was not observed in any other run. Similarly, run ten is the only run showing three different classes of active social

learners without turn-taking, with varying levels of activity (i.e. '7-ASL without...', '7a-more active SL without...', and '7aa-even more active SL without...').

However these classes were not observed consistently across all the runs and their presence or absence might be a by-product of the changes in the learning design or affordances of the platform itself. Examples of such, is the occurrence of the significantly-sized clusters '3-initiators without replying', and '3a-more active initiators without replying', which were only observed in the earlier runs of Understanding Language, but was only observed once again, in run 11, and in a much reduced number. This is similar to the large clusters of '4-initiators who respond' which appeared in early run of both MOOCs but in this case are never seen again. In fact, the demise of the initiators' classes coincides with the emergence of the new classes '7a-more active SL without turn-taking' and '8a-more active social learners'. It is quite possible that learners in these classes migrated from a low-engagement class to a higher-engagement class, following the introduction of the intervention by FutureLearn mentioned in Section 5.3, by which learners receive email notifications when others reply to their comments. This intervention took place in early 2016, between runs three and four in both MOOCs, as per the dates shown in Table 5.1.

Another observation to make from observing the barcharts in Figures 5.15 and 5.16 is that there are some categories defined by Chua et al. (2017) which rarely appear in a MOOC (if ever) and when they do, it might be with a very small numbers of learners. This is the case of '5-repliers', which appears in Portus but not in Understanding Language, and '6-reluctant ASL' which appears only in a few runs of each MOOC. In particular, there are only five, six, and two repliers in the second, third, and fifth runs of Portus respectively.

### 5.7.5 Comparing against the learner types as per Chua's heuristic

Despite a carefully-chosen naming convention for the labels to order them in the  $x$  axis, further insights from the charts in the previous section are still hard to extract. This is due to the results being disaggregated by the categories found by the clusterer which, as discussed, do not appear consistently across all runs of MOOC. Adding the clusters containing instances of specialisations from the original categories in each run results in the distributions in Figures 5.15 and 5.16. These show the barcharts of categories per run when these have been aggregated according to the names of the classes as per the heuristic, e.g. adding '2-loners' to '2a-more active loners' in a given run, and all variations of '8-active social learners' together (i.e. '8-active social learners', '8a-more active social learners', '8aa-even more active social learners', '8b-ASL who do not give

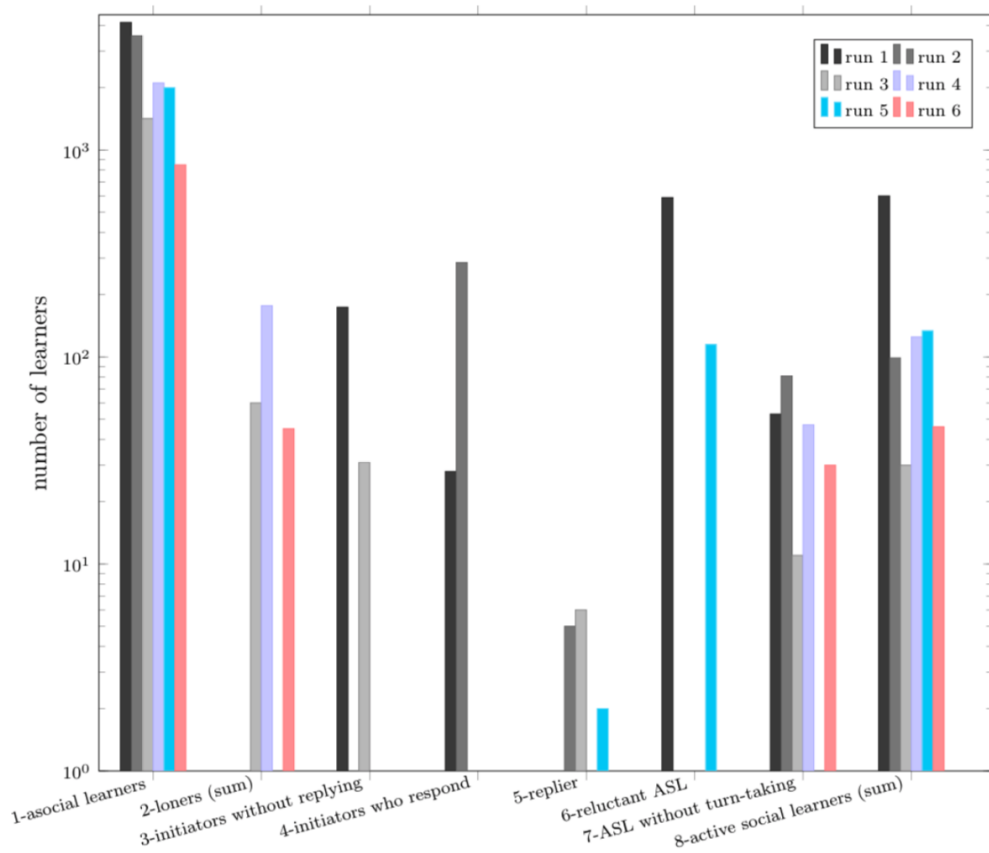


FIGURE 5.15: Distribution of social learners across on Portus according to the clusters found by X-Means and aggregated guided by Chua's classification

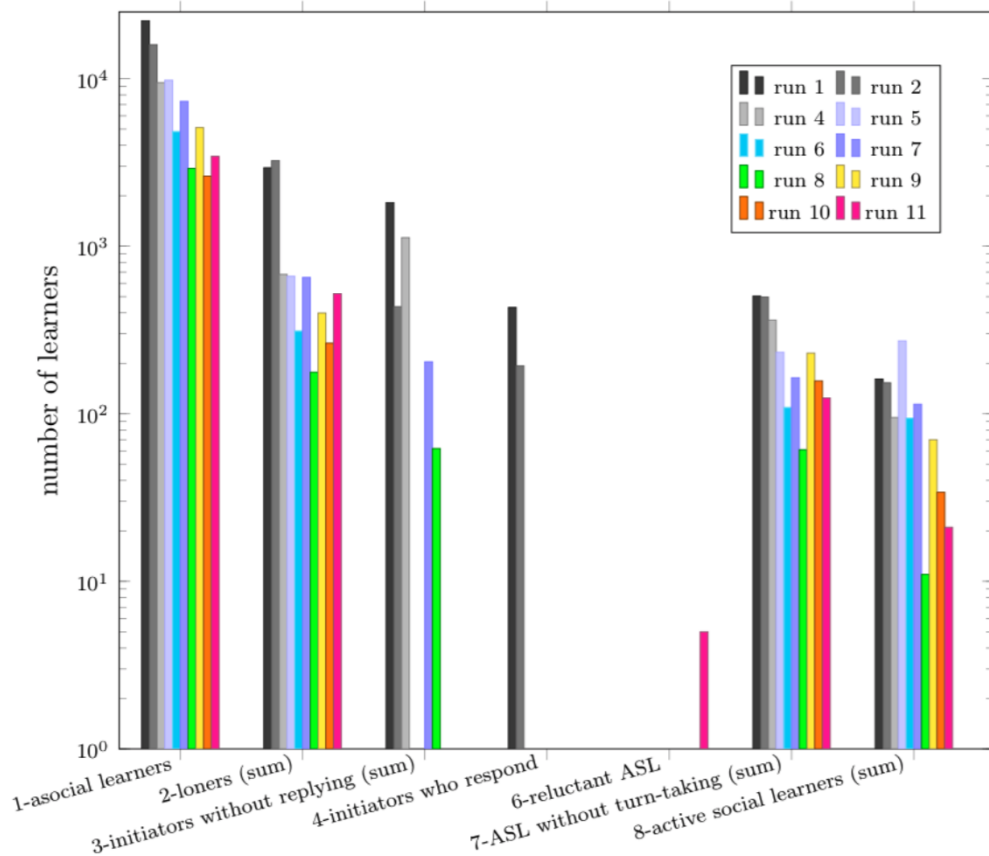


FIGURE 5.16: Distribution of social learners across on Understanding Language according to the clusters found by X-Means and aggregated guided by Chua's classification

additional replies’, and ‘8bb-more active SL who do not give additional replies’).

This aggregation allows for an important like-for-like comparison between the classes found via a data-driven approach (since clustering is an unsupervised learning algorithm) and those set by the application of the heuristic by [Chua et al. \(2017\)](#) with the addition of the asocial class. This entails inspecting these figures and comparing them against Figures 5.2 and 5.3, which showed the distributions of learners in each of the runs for the MOOCs according to the heuristic explained.

At first glance there seems to be a correspondence between the clusters found via unsupervised learning, and those defined by the heuristics. However, some important differences are noticeable, beyond the minor issue of the labels not being in the same order (as the chosen ordering served different purposes each time). The most recent, Figures 5.15 and 5.16, show much fewer learners in some of the classes (as discussed in Section 5.7.4), but also some other classes with much more learners than their near-homonyms in the heuristic case. This insight can be made more precise through comparing the exact numbers in Tables 5.4 and 5.13 side by side for Portus, and those in Table 5.15. For example, there are many more ‘1-asocial learners’ than ‘asocial learners’ in each of the runs of both MOOCs<sup>23</sup>.

Therefore it is clear that the apparent correspondence between the heuristic and the data-driven approach is not quite an exact match, and that many of the learners considered by the heuristic as “social” (because they might have made a comment of certain kind) are in fact considered by the clusterer as exhibiting behaviours much closer to that of an asocial learner than to any of the other categories that the nature of their given comment would have placed them into. Further, this “migration” of learners from heuristic-based classes into data-driven classification is much more widespread than the few examples given in Footnote 23, as Figure 5.17 evidences. To create this figure I constructed the matrix so that the “ground truth” were the clusters generated by X-means (with  $k = 7$ ) and the “predicted class” were those determined by the heuristic. The clusters were renamed in order to pivot the columns so that, if the match of clusters to classes was good, it would appear largely as a diagonal, with a high correspondence between clusters and classes.

This matrix shows that all asocial learners (in cluster6, now c1) were correctly identified as ‘1-asocial learners’, as the top left value for each matrix are exactly the same values as those in the corresponding entries in Table 5.4 show for all runs of Portus

---

<sup>23</sup>In Portus 1 there were 3,261 asocial learners, yet the X-Means algorithm identified 4,134. Portus 2 saw an increase from 2,748 asocial learners into 3,562 ‘1-asocial learners’. For Portus 3, the increase was from 1,205 to 1,418; for Portus 4 it was from 1,716 to 2,107; for Portus 5 from 1,704 to 2,000; and for Portus 6 from 751 to 848. The same phenomenon was observed in Understanding Language, across all of the ten runs I examined.

DIAL\_portus\_all\_runs\_7clusters.csv

1-asocial	11385	0	0	0	0	0	0
2-loner	2246	150	16	1	0	0	0
3-initiator without replying	465	196	40	7	0	1	2
4-initiator who responds	111	87	19	8	0	1	0
5-replier	27	1	0	0	4	0	0
6-reluctant asl	38	19	3	1	1	0	0
7-asl without turn-taking	15	4	2	1	2	0	0
8-active social learner	315	651	233	96	18	30	148
	c1	c2	c3	c4	c5	c6	c7

FIGURE 5.17: Confusion matrix plot for the matching of clusters found by the X-Means clustering algorithm on the Portus MOOC (all runs), with  $k = 7$ , against those categories identified by applying Chua's heuristic. The mapping for cluster names that allow pivoting the confusion matrix (for improved readability) is as follows: cluster1→c2, cluster2→c7, cluster3→c6, cluster4→c4, cluster5→c3, cluster6→c1 and cluster6→c5.

(11,385), meaning that the recall for this class was 100%. However, the precision is poor as many instances across the other categories in the heuristic are also deemed to be '1-asocial'. This is due to the strictness of the definition of the heuristic in [Chua et al. \(2017\)](#), whereby a learner is deemed to be of a certain class by having posted at least one comment of a given type. For example, someone having produced just one lone post would be deemed a 'loner' whilst someone having given only one reply would be a replier, despite neither of them having had any more engagement throughout the course. The clusterer would rightfully deem these two cases as having more in common with each other (and with asocial learners) than with loners or repliers who may have produced many more posts of each kind. It is clear that the intensity of interactions is a very important differentiator of learner groups, as much as the type of these interactions is. Whilst for [Chua et al. \(2017\)](#) only the type of interactions is taken into account, for much of learning analytic research I reviewed it is only the intensity of interactions which guides the classifications ([Milligan et al. \(2013\)](#); [Kizilcec et al. \(2013\)](#); [Kizilcec and Chen \(2020\)](#); [Alario-Hoyos et al. \(2014\)](#); [Anderson et al. \(2014\)](#); [Ferguson and Clow \(2015a,b\)](#), amongst others). A clustering approach such as the one presented in this thesis, which combines both type and intensity of interactions, is an important contribution to this space.

## 5.8 Summary and conclusion for this chapter

The feature engineering process described in Chapter 3, informed by the model defined in Chapter 4 on the data described in Section 5.2 produced an anonymised dataset with 78 features extracted on learner engagement data in sixteen FutureLearn MOOCs, comprising dialogic features, interval features, and badge features amongst others. An analysis of the performance of a clustering algorithm on three feature subsets led to the selection of dialogic features. Through a similar performance analysis, the clustering algorithm X-Means was selected as it was superior by far amongst those trialled. Less clear were the performance gains in selecting a number of clusters amongst experimentation between two and ten. In particular, the choices of  $k = 4$  and  $k = 7$  seemed equally appropriate from this perspective, but I settled for seven clusters as this was closer to the number of classes in the heuristic by [Chua et al. \(2017\)](#).

In the sixteen experiments involving each run of each MOOC, all of these heuristic classes were observed in the clusters (as well as several additional variations, to account for intensity of interaction or further specialisation of a class). However, not all the classes were observed, and only four were consistently found across most of the runs, as shown in Table 5.16. This suggests that having chosen  $k = 4$  instead of seven would have probably led to similar conclusions overall.

TABLE 5.16: Summary comparative table of clusters.

Categories found in most runs	Clusters found with X-Means ( $k = 7$ ) in at least one run of Portus or Understanding Language	Social learner groups in the heuristic by <a href="#">Chua et al. (2017)</a>
Asocial learners	1-asocial learners	N/A
'Loners'	2-loners	Loners
	2a-more active loners	
(Active) social learners without turn-taking	3-initiators without replying	Initiators without replying
	3a-more active initiators without replying	
	4-initiators who respond	Initiators who respond
	5-replier	Repliers
	6-reluctant ASL	Reluctant active social learners
	7-ASL without turn-taking	Active social learners without turn-taking
	7a-more active SL without turn-taking	
	7aa-even more active SL without turn-taking	
Active social learners	8-active social learners	Active social learners
	8a-more active social learners	
	8aa-even more active social learners	
	8b-ASL who do not give additional replies	
	8bb-more active SL who do not give additional replies	



The categories more commonly found across most runs of both MOOCs are:

- A cluster of *asocial* learners, who did not make any comments, or had exhibited so little activity that it was deemed by the algorithm to be more similar to learners in this class than to any other.
- A cluster of *'loners'*, who despite having posted comments, these tended not to spark any comments from peers.
- A cluster of *(active) social learners without turn-taking*, including those who tended to initiate conversations (some, if not all, their posts were commented upon), and also replied to others. I place the word “active” in parenthesis, to denote that the key term is not the level of activity but rather that they tended to miss out on the turn-taking nature of conversations with peers.
- A cluster of *active social learners*, comprising those who, in addition to engaging in the behaviours by the other groups, would also reply under their own initiating posts and do additional replies.

In addition to the above, other clusters of nuanced behaviours were observed in each of the runs, roughly falling in the above categories but with sufficiently distinct data characteristics to be picked up by the X-Means algorithm as separate clusters.

The analysis of the results of such experiments provide answers to the research question **RQ2** of this thesis: *What does a data-driven approach to learner interactions reveal about learning engagement within FutureLearn MOOCs?* In summary, it reveals that:

- unsupervised learning algorithms such as clustering in general, and X-Means in particular, are useful to discriminate different classes of learner interaction behaviours naturally occurring in FutureLearn MOOCs;
- learning design changes impact posting behaviour of learners, with simple nudges such as email notifications being possibly responsible for the disappearance (or significant reduction) of behaviour classes such as ‘initiators without replying’ and the emergence of higher-activity classes such as ‘more active social learners who do not give additional replies’, amongst others;
- learners’ posting behaviour in MOOCs follow the 90-9-1 rule, where the large majority of learners ‘lurk’ and do not produce any posts (hence called ‘asocial’ here), with few learners being very active participants, producing hundreds of posts each; and,

- dialogic heuristics such as that by [Chua et al. \(2017\)](#) are helpful in defining a nomenclature in the analysis of clusters of learner engagement in conversation though are not sufficient to capture nuanced behaviours defined by the intensity of interactions in a given course which emerges from data-driven approaches.

## Peer-learning in face-to-face instruction mediated by PeerWise

*“The relationship between the teaching and research is the same as between the confession and sin: If you have not sinned, then you have nothing to confess!”*

Anonymous, quoted in “HOW TO SOLVE IT: MODERN HEURISTICS”, by Z. Michalewicz and D. B. Fogel, Springer.

Though the epigraph above suggests that the ability to teach a topic is gained through having researched it, certainly much research can also be born out of teaching. Indeed, much of the literature discussed in Section 2.3, if not all, was on research emerging from the use of PeerWise in teaching practice. This is also the case for the research I present in this chapter, which was motivated by observations made whilst lecturing in a second-year module in Human-Computer Interaction at the University of Southampton, where I was responsible for the assessment design and learning activities of two cohorts in consecutive years.

In particular, the lens I use in this chapter to analyse the engagement of students in these two cohorts is research question **RQ3**: *What does a data-driven approach to learner interactions reveal about learning engagement within the PeerWise digital environment for face-to-face instruction?* To answer this question, I applied the model presented in Chapter 4 to data collected through a quasi-experiment<sup>1</sup> on two consecutive offerings of the same module, the motivation and context of which is detailed in Section 6.1.

<sup>1</sup>This is a quasi-experiment as defined in Section 2.1.2, as it is an ex post facto study on an opportunistic sample of two non-random groups of students exposed to one variation in the learning design.

The datasets related to each offering of the module are described in Section 6.2. Section 6.3 gives details on the application of the model on this platform to engineer suitable features as described in the methodology in Chapter 3. The feature extraction process on these datasets is given in Section 6.4, followed by the application of the clustering algorithm on the data using the engineered features, in Section 6.5.

The results of the clustering allows for a direct comparison against the findings on MOOC engagement as shown in the previous chapter. Such a comparison would provide an answer to research question **RQ4** *Is learner engagement different in different kinds of peer-supported digital environments, be it a complement to face-to-face instruction, or a fully online course?* This is presented in Section 6.6. A summary and conclusion for this chapter is given in Section 6.7.

## 6.1 Motivation and context

Recent years have seen an increase in research interest in learner engagement within peer-supported digital environments. Yet, within said spaces, the effect of lecturers' participation incentives on the quality of learner interactions has been little explored. This chapter presents one such study of learner engagement over two consecutive years of using the web-based peer-learning software PeerWise.

This study lies at the intersection of two separate but related trends in higher education: Firstly, multiple-choice questions (MCQs) becoming more pervasive in student assessment, due to a combination of larger student numbers, reduced teaching resources and the greater efficiency afforded by computerised MCQ marking (Nicol, 2007). This trend towards more MCQ-weighted assessment has however been outlined as problematic, particularly so in discursive subjects, such as those from the arts and humanities, given that previous findings about MCQs being associated with promoting memorization and recall to the neglect of higher order cognitive processes (Scouller, 1998; Nicol, 2007). The second trend is the increased interest in the study of learner data from peer-supported environments such as Course Management Systems (CMSs) and Virtual Learning Environments (VLEs), as mentioned in Section 2.4. The rising interest in learning analytics applied to data from these environments is typically focused around issues of predicting attainment (much like learning analytics in MOOCs much too often focuses on predicting dropout). In many of these environments however, there are conversational capabilities that can be studied with learning analytics for a more complete picture about engagement, which has not been done amongst the reviewed literature thus far, other than those following Social Network Analysis approaches (as cited in

Romero and Ventura (2010)).

Having outlined the wider context of this part of my research, I now move on to describe the specific context within which it is situated, including some details about my previous research<sup>2</sup> in this space, much like I did in Section 5.1 in MOOCs.

The context of application was the 12-week long module *Interaction Design* (code COMP 2213), a second year module at the University of Southampton on topics of human-computer interaction. This is a compulsory module for Computer Science students at the University of Southampton, and optional in several other courses, including Psychology and Web Science. As described in Wilde and Snow (2018a), faced with increasing student numbers (from less than 80 to over 160 in two years<sup>3</sup>), I worked on redesigning the assessment with the main intention of providing, with other lecturers<sup>4</sup>, timely feedback to challenging, engaging coursework. Particularly challenging for this module is the fact that much of the content of the module is discursive in nature, as it introduces knowledge and skills from social sciences and arts (including design thinking, design theory and qualitative methods) with an emphasis on the human elements of computing (including cognitive psychology). This is inherently different from students' previous experiences in computer science modules.

Introducing PeerWise to the module was amongst the changes to the learning design described in Wilde and Snow (2018a). With some colleagues in my teaching team, I observed that the first cohort had used PeerWise as a revision aid beyond the requirements of the module, as described in Snow et al. (2018). There, details on the experiences of students' and lecturers using PeerWise for first time for Interaction Design are given. With my co-authors analysed learner engagement in this module and presented results of a mixed-methods analysis of learner engagement with the software (Snow et al., 2018). In particular, we used qualitative methods to explore the themes that students reflected upon, around issues of use of software as a revision aid, appropriateness of use for the learning matter, and affordances of collaboration amongst several others. In terms of quantitative methods, we used descriptive statistics to identify exam marks distributions amongst those students who used the software within 48 hours before the exam, the numbers of questions and answers given over time (noting the intensity of the engagement increasing around deadlines) and the proportion of correctly answered questions given the number of submitted answers per question amongst others.

---

<sup>2</sup>Some of the findings in this chapter have been published: a comparison of the two cohorts (Wilde, 2020), and work focused on the first cohort (Wilde, 2019; Snow, Wilde, Denny, and m.c. schraefel, 2018; Wilde and Snow, 2018a; Snow and Wilde, 2017). This chapter reports my contributions only.

<sup>3</sup>The doubling of student numbers, referred to above, concerns the academic years between 2014/15 to 2016/2017. Nowadays the number of students enrolled in this module is around 300.

<sup>4</sup>Also in the teaching team were: m.c. schraefel (only in 2015/16) Enrico Costanza (also only in 2015/16), Nick Gibbins (only in 2016/17) and Steve Snow (both years).

In [Wilde \(2019\)](#) I expanded upon the results published previously by highlighting the differences in observed behaviour on the second cohort using PeerWise for this module which had not been yet studied in any significant depth, and offered a strategy for doing so. Results on a comparative statistical analysis of both cohorts were presented in [Wilde \(2020\)](#), and the steps informing the feature engineering conducted for unsupervised learning for these datasets, in [Wilde \(2021\)](#).

## 6.2 Datasets and learning design

In addition to the data used in the analyses summarised above and presented at length in [Snow et al. \(2018\)](#), there are data related to the use of PeerWise in this module for a second year, having however first made a variation to the learning design. In this section I give details that are common to both cohorts and what were the differences in the datasets available for analysis.

For both cohorts, the module was assessed with a computer-based exam at the end of term (on *QuestionMark* perception) worth 50% of the marks of the module, which included just over twenty multiple choice questions, six short answer questions, three fill-in-the-blanks questions, as well as two longer-answer questions. All other assessment was assignment-based. The main assignment, common to both offerings of the module, was the creation of a low-fidelity prototype for an Internet of Things application with a report and a video as deliverables<sup>5</sup>.

At a high level, participation data can be described as characterising learners from two semester-long deployment of PeerWise, with two classes of 141 and 169 Computer Science students, in 2015/16 and 2016/17 respectively, who authored and answered multiple-choice questions (MCQs) in Interaction Design, in topics such as cognition, requirement elicitation and prototyping. PeerWise was enthusiastically adopted as a tool for exam revision, given that it had a large element of assessment via MCQs within. For each cohort, a ‘course’ in PeerWise was created (courses 12710 and 14715) and the dataset structure associated to each of these is shown in [Figure 3.5](#). Additionally, for each cohort there were assessment datasets, the structure of which was shown in [Figure 3.6](#).

I obtained overarching ethics approval<sup>6</sup> to use the datasets regarding students’ participation through the PeerWise software which was used for students’ authoring and

---

<sup>5</sup>More details on how video was used for assessment in human-computer interaction are available in [Wilde and Snow \(2018a,b\)](#); [Vasilchenko et al. \(2018\)](#); [Wilde et al. \(2019\)](#); [Wilde \(2019\)](#); [Wilde and Dix \(2020a,b\)](#).

<sup>6</sup>ERGO FEPS 55694, as shown in [Appendix A](#).

answering Multiple-Choice Questions (MCQ) for both cohorts of this module.

### **6.2.1 The first cohort: class of 2015/16**

PeerWise was first deployed within the module Interaction Design in the second semester of 2015/16 (January-May 2016). Aside from the more discursive flavour of Interaction Design compared to the science and maths-based modules to which PeerWise is typically deployed (as seen in Section 2.3), the implementation in Southampton was similar to these deployments in the literature. Students were required to author four questions (and answer four others) over the course of the semester. This compares to author one and answer twenty (Denny, 2013), author four and answer twenty (Renzo et al., 2014), and author one and answer one per teaching week (i.e. in a semester approximately author and answer thirteen questions, (McClean, 2015)). Five percent of the module's total marks were allocated to participation in PeerWise, a practice which is consistent with what is reported in the literature – i.e. 3% (Bates, Galloway, and McBride, 2012), 10% (Devon et al., 2012) and 1.5% (Denny, 2013). A further 5% was awarded to a reflective essay about the use of PeerWise in class and how the delivery could have better enhanced their learning.

The class of 140 students was divided into 27 groups of 4–6 students each. Students worked in these groups throughout the course on all non-exam assessments. The PeerWise component accounted for 10% of the total course marks. This involved both participation in PeerWise (5%) and the reflective essay (5%). However, neither questions submitted by students nor their answers were awarded formal marks.

To achieve the full participation mark, each group had to ensure that every member of their group (1) authored at least four questions, and (2) answered at least four questions by the final deadline of April 26<sup>th</sup>. At least one of these questions and answers was required by a mid-semester deadline on March 18<sup>th</sup>. Though question content was not assessed students were warned that any irrelevant or nonsensical questions would be removed and any bullying or offensive language would be penalised. Students were asked to comment on each question they answered, but this was not policed. It was requested that the content of the questions authored prior to the first deadline in March should reflect content covered in class up to that point and subsequent questions should cover course content from this deadline, up to the final April deadline. An additional requirement was for each question to include the group number in the title of the question. This was partly for accountability, so that although students individually remained anonymous, questions could be traced to individual groups, and partly for attributing the participation mark.

For the written reflection mark (5%), each group was required to co-author a 1,000-word reflective essay on their experience with PeerWise in supporting their learning in the module. The results of the qualitative analysis of these essays are presented in [Snow et al. \(2018\)](#).

The important point for the context of this Thesis is that for the first cohort (in 2016) students were required to use PeerWise in a compulsory manner. As a motivation for the use of this software, students were told the following as part of the coursework specification (available in full as Appendix F):

*“For this coursework, you are required to formulate multiple choice questions (MCQs) on topics in Interaction Design to aid your exam revision. This would enable you to have a good understanding of the examiners’ likely frame of mind when producing questions on the examinable content (McMillan & Weyers, 2011). In addition, you will be supporting each other’s learning by answering questions formulated by your peers, and offering your feedback, demonstrating your comprehension of the materials covered in this module. To support this work, the online platform PeerWise (<https://peerwise.cs.auckland.ac.nz>) is used. Instructions for registering can be found in the NotesWiki.”*

By including this message in the coursework specification, a further nudge for participation was given as an encouragement that their authoring and answering of MCQs might be beneficial in their exam preparation (given that half of the exam marks were in the form of MCQs.) This, added to the participation incentives given through the assessment design (whereby each member of the group had to meet the individual participation threshold to be eligible for the 5% participation marks), formed a strong behavioural constraint towards the use of this tool for this cohort.

From the results presented in [Snow et al. \(2018\)](#), the following observations are particularly relevant for this thesis:

1. With the exception of a very active minority, the majority of students in the first cohort took the minimum effort approach, i.e. submitted only the minimum number of questions and answers required by the deadline, with additional answers submitted for revision prior to the exam. For the second cohort there was even less participation by the majority, as there were no set deadlines or incentives to participate.
2. The majority of questions authored by students in the first cohort were submitted directly prior to the first deadline and thus the final PeerWise question-bank at the end of semester was skewed towards content covered earlier on in the module.



3. Despite this, PeerWise was used extensively as a tool for exam revision, particularly so by a dedicated few.

The first two observations from the list above seem to suggest that the learning design constituted a strong behavioural constraint for the majority of students, as they were ‘coerced’ to participate (for lack of a better word) or else they would be eligible to those marks. These students met the strict requirement and engaged no further (indeed typically doing so just before the participation deadlines that were imposed by the lecturers). It is therefore evident from this engagement data that the removal of such constraints from the learning design would allow learners to behave more freely. However at this stage it was not clear, for these learners, what that freedom would have looked like.

Finally, the fact that there were only a “very dedicated few”, as per the third of these observations became even more evident in the second year, as discussed in Section 6.2.2.

### **Summary statistics of the participation in PeerWise**

Of the 140 students in the class, 132 contributed to PeerWise one or more times during the semester. As many as 531 separate questions were authored with 8,679 questions answered and 312 comments made. Notably, more than half of all questions authored were submitted shortly prior to the first deadline, by when each group member were required to have submitted at least one question and one answer for the group to be eligible for full marks in that part of the assessment. A further 167 questions were authored after the first and before the second deadline (Snow et al., 2018).

## **6.2.2 Learning design changes for the second cohort: class of 2016/17**

The use of PeerWise was mandatory for the first cohort, with marks awarded subject to a minimum level of participation by two given deadlines. In contrast, for the second cohort this was an optional activity, not rewarded with marks, as shown in Table 6.1 which details the differences in assessment in this module between these two consecutive academic years.

The removal of the incentive to participate resulted in a lower uptake in the second cohort, which in turn led to far fewer questions being produced, as shown in Table 6.2 (alongside other comparative statistics and relevant details about these cohorts). Despite the lower number of student-authored questions, these were judged by peers to

TABLE 6.1: Differences in assessment design between the two offerings of Interaction Design under study. Note that there are marks given to participation in PeerWise in the first deployment, which is absent in the second.

	<b>First cohort</b>	<b>Second cohort</b>
Academic Year	2015/16	2016/17
Assessment design	Computer-based final exam with an MCQ part (50%) Other coursework, including video coursework (35%) PeerWise participation (5%) PeerWise reflective report (5%)	Computer-based final exam with an MCQ part (50%) Video coursework with report (50%)

be of much higher quality overall. In fact, both the quality and the difficulty of the student-authored questions were perceived as significantly higher in the group with optional participation in PeerWise, with difficulty averages of 0.558 and 0.722, and quality averages of 2.522 and 3.037 respectively, as shown. Other metrics for engagement in this peer-supported environment were comparatively high, in particular, those related to “conversations” sparked from questions. These metrics incorporate replies to comments added to questions, and the actors involved in the exchanges, as calculated using the model of learner engagement in peer-supported digital environments, presented in Chapter 4.

Particularly noteworthy is the fact that even though the number of enrolled learners in the second cohort is larger (169 from a previous 141), there were only 107 students who used the platform (only 62% of the total for the year). This is in stark contrast with a near-total adoption in the previous year when 139 students were active amongst the 141 who had enrolled (over 95% of the registered students complied and engaged with the software).

Students from the second cohort created collectively only 81 questions, compared to the 531 generated by the previous cohort, due to a much lower number of unique authors (22 against 126), and a slightly lower average production effort (three questions per author against 4.2 in the whole semester). However, the minimum requirement to achieve full participation marks was the production of four questions. Other metrics of cohort participation were also lower, such as far fewer comments (265 against 118) and answers to questions (8,707 against 4,993).

A discussion on these observations on the comparative statistics of these two deliveries of Interaction Design was given in [Wilde \(2020\)](#), where the interesting tension between wider adoption and quality of the interactions was pointed out, as well as other

perceived benefits of the participation within PeerWise: all of the 107 students who used it in the second year achieved exam marks of 50% or higher, of which, 99 obtained 60% or higher. A limitation of this study is that it does not control for self-selection bias, i.e. I cannot claim that the use of PeerWise caused students to succeed in the exam (it is possible that only “good” students chose to use the software). It is interesting to note, however, that the assessment design does seem to affect student engagement even if learning activities and content remain the same.

The descriptive statistics on these datasets suggest that a learning design intervention, such as awarding marks as an incentive for wider participation, encourages wider participation yet might disincentivise deeper connections. Conversely, removing this incentive for participation may foster a higher quality of content and of interactions, which calls for a deeper analysis on the effect of the intervention on the learner engagement in this particular platform.

TABLE 6.2: Comparative statistics and other characteristics of the consecutive offerings of the Interaction Design module.

<b>Characteristics of each deployment</b>	<b>First</b>	<b>Second</b>
Academic Year	2015/16	2016/17
PeerWise Course ID	12710	14715
Students enrolled	141	169
PeerWise active learners	139	107
Lecturers in the module	4	3
Group size for coursework	4-6	4-6
Questions authored	531	81
Answers given	8,707	4,993
Comments made	265	118
Replies given	45	92
Number of ratings	4,775	2,530
Average ratings given	0.558	0.722
Average quality of question	2.52	3.04
Followers	30	27

### Datasets details

The descriptive statistics for both cohorts of Interaction Design who engaged with this software, discussed above, were performed on the PeerWise datasets here listed, which follow the schema shown in Figure 3.5. The shape of each of these files in comma-separated values format (CSV), which include the file headers, is shown in Tables 6.3

and 6.4. Hence, as for example, the file `Users_12710.csv` has 139 rows and three columns, it contains information about 138 learners in the first cohort, specifically, their PeerWise `User_ID`, their `Username` and their `Identifier`, as shown in the schema in Figure 3.5.

In addition to the PeerWise participation data, I incorporated data of the students' attainment, both in the QuestionMark Perception exam (which had a 50% element of MCQ) and the overall marks in the module (which include the coursework element). These are reflected in the Grades CSV. Note that there are three more columns in `Grades_12710` than in `Grades_14715`, reflecting a difference in the learning design.

TABLE 6.3: Files in the 2015/16 dataset for PeerWise. All the listed files have extension `.csv`.

File name	rows	columns
<code>Users_12710</code>	139	3
<code>Questions_12710</code>	532	16
<code>Comments_12710</code>	266	5
<code>Replies_12710</code>	46	6
<code>Followers_12710</code>	31	3
<code>Ratings_12710</code>	4776	6
<code>Badges_12710</code>	138	26
<code>Answers_12710</code>	8708	5
<code>Groups_12710</code>	145	3
<code>Grades_12710</code>	135	27

TABLE 6.4: Files in the 2016/17 dataset for PeerWise. All the listed files have extension `.csv`.

File name	rows	columns
<code>Users_14715</code>	108	3
<code>Questions_14715</code>	82	16
<code>Comments_14715</code>	119	5
<code>Replies_14715</code>	93	6
<code>Followers_14715</code>	28	3
<code>Ratings_14715</code>	2530	6
<code>Badges_14715</code>	106	26
<code>Answers_14715</code>	4994	5
<code>Groups_14715</code>	171	3
<code>Grades_14715</code>	169	24

## 6.3 Modelling interactions in PeerWise

In this section I explain how to apply the model presented in Chapter 4 to describe the various relationships and interactions between users of PeerWise. To that aim, I will use a motivating synthetic example to describe the kind of data collected when students interact within the platform, both with each other, and with the subject matter, as stored in the datasets with in the schema presented in Figures 3.5 and 3.6.

Let us consider the following example as shown in Figure 6.1. In this simplified scenario, let there be three students  $s_1$ ,  $s_2$ ,  $s_3$ , where each student creates a number of multiple-choice questions, represented as  $q_{i,j}$ , with  $i$  representing its author, and  $j$  indicating it refers to the  $j$ -th question created by student  $i$ . Thus, student  $s_1$  authored two questions ( $q_{1,1}$  and  $q_{1,2}$ ), student  $s_2$  authored four questions, ( $q_{2,1}$ ,  $q_{2,2}$ ,  $q_{2,3}$  and  $q_{2,4}$ ), and student  $s_3$  authored only one question ( $q_{3,1}$ ).

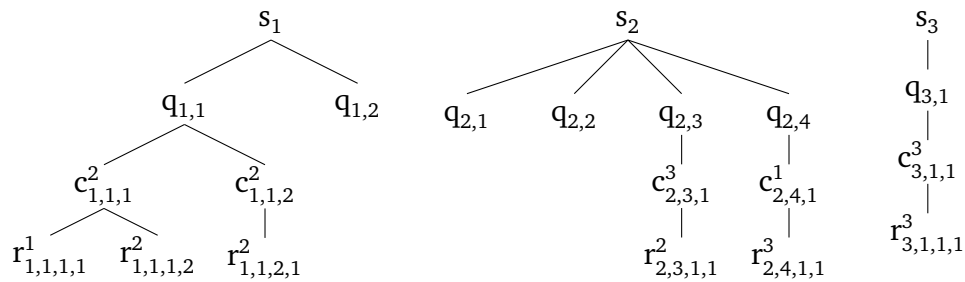


FIGURE 6.1: A simplified test case with three students in PeerWise, creating questions, comments and replies. In this example, student  $s_i$  makes question  $q_{i,j}$ , which in turn raises comment  $c_{i,j,k}^t$  by learner  $s_t$ , and student  $s_p$  gives a reply  $r_{i,j,k,l}^p$  to the comment.

The authorship relation is shown in the respective three trees in Figure 6.1 by the edges connecting any  $s_i$  student with a question  $q_{i,j}$ , with each tree being rooted by a student  $s_i$ . This relationship is reflected in the PeerWise dataset in the related *Questions* file for the course (in this case, *Questions\_test*, shown in Table E.3, in Appendix E<sup>7</sup>). Note that the relationship ‘question-author’ is captured in the file by associating to each question a unique key (see column *ID* in Table E.3) and the author’s *Identifier*.

<sup>7</sup>Appendix E provides tables listing the relevant content to the CSV files for the synthetic dataset for the example in Figure 6.1, namely *Users\_test*, *Questions\_test*, *Comments\_test*, *Replies\_test*, and *Answers\_test*. These were created in order to test that the feature extraction process followed the model of learner interactions described in Chapter 4.

Another observation to make about the reduced example is that it has the same structure as the trees shown in Figures 4.6 and 4.9, that modelled the interactions between the fictional learners Ana, Bob and Cam. In this case, however, rather than comments in a chat (or a discussion thread of a MOOC), the trees represent the authoring of MCQs within PeerWise and the conversations amongst peers who have been exposed to them. As seen, once a learner submits an MCQ to the question bank of their course in PeerWise, it allows others not just to answer them (and test their knowledge on the content matter of the question), but also to *critique* them. This is possible to do by giving the question a rating and a difficulty score, as seen in the comparative statistics in Table 6.2. Most importantly, however, from a communicative perspective, this critique can be done through dialogue.

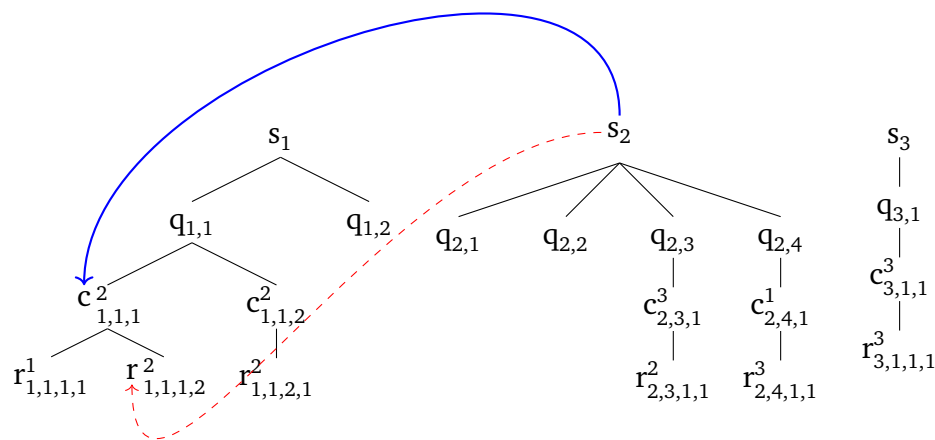


FIGURE 6.2: Test case graph of students, questions, comments and replies as per Tables E.4 and E.5. The supra-index notation for comments and replies allows to keep track of comments' and replies' authors. As an example, the additional lines show that student  $s_2$  authored comment  $c_{1,1,1}^2$  (the solid blue line) and reply  $r_{1,1,2}^2$  (the red dashed line)

Therefore, though question making would ordinarily be a non-communicative activity (as the creative “interaction” is between the individual learner and the content matter), through the platform affordances it becomes a communicative activity, since others can critique them. The digital traces of the process, as preserved in the relevant files represented in Appendix D are the e-tivities from the model.

To use the vocabulary given by the model, all questions  $q_{i,j}$  are then zero-order replies, which may or not spark conversations. When they do, they are *starting questions* in PeerWise (or SP in the model), such as  $q_{1,1}$ , which sparks comments  $c_{1,1,1}^2$  and  $c_{1,1,2}^2$ . When they do not, they are *lone questions* in PeerWise (or LP in the model), like  $q_{1,2}$ .

For this synthetic example, the index associated to the student who created a comment is shown as the supra-index of the comment, such as in the case of  $c_{1,1,1}^2$ , which

was authored by student  $s_2$ , as emphasised in Figure 6.2 with a blue arrow making the authorship connection which would not ordinarily be evident at first glance. In the motivating example with Ana, Bob and Cam, presented in Chapter 4, the authorship of comments and replies was made evident through choosing a different colour for each learner (Ana in pink, Bob in gray and Cam in teal, in Figure 4.6). In the formal model, this is captured with the second element of the tuple  $\langle \text{what}, \text{who}, \text{when} \rangle$ , as shown in Figure 4.9. In the implementation for the PeerWise platform, all of these elements are captured in the columns of the `comments.csv` file, as seen in the schema in Figure 3.5. Once again, to use the vocabulary given by the model, all comments  $c_{i,j,k}^p$  are then first-order replies, unless  $p = i$ , as in the case of  $c_{3,1,1}^3$ , which is a comment made by student  $s_3$  on their own question  $q_{3,1}$ , which is according to the model, an LP as it is not part of a conversation with others.

Finally, only the second-order replies remain to be discussed, and these are in PeerWise the replies given to existing comments. When the reply is made by the original author of the question, such as  $r_{1,1,1,1}^1$ , given by  $s_1$  in the example, it is an *initiators' reply*, and when it is by someone else, it is a *further reply*, such as  $r_{1,1,1,2}^2$ , given by  $s_2$ , as emphasised in Figure 6.2 with a red dashed line.

## 6.4 Features on PeerWise data

This section presents the features that are either directly extracted or engineered from the datasets about the learners in Interaction Design. Table 6.5 and Table 6.6 show the features extracted from the PeerWise dataset and the assessment files.

In particular, the features shown in Table 6.5 are, in addition to the unique identifier for the user (anonymous and given by PeerWise), all of the badges offered within PeerWise that were listed in Table 2.6. Amongst them, it is worth making some additional observations regarding  $B1$ ,  $B4$ , and  $B5$  in particular, as these three features inspired the “badges” features I engineered in MOOCs (as explained in Section 5.6).  $B1$  is ‘True’ if the learner has authored at least a question, which would be equivalent to having engaged in at least one zero-order reply in the model (it will be true if SP or LP are greater than zero). Similarly,  $B4$  is ‘True’ if the learner has written at least one comment under any question, so at least one first-order reply (and it will be true if FR is greater than zero). Finally,  $B5$  is ‘True’ if the learner has given at least one reply to a comment written about their own question, that is, an initiator’s reply, which is a specific kind of second-order reply (the other one being additional reply). It will be true if IR is greater than zero.

TABLE 6.5: Features extracted from the PeerWise dataset

Feature	Type	Alternative name	Description
<i>User_ID</i>	string		unique identifier in PeerWise for a student <i>s</i>
MILESTONE BADGES (CAN BE EARNED ONLY ONCE)			
<i>B1</i>	numeric	Question author	<i>s</i> contributed one question
<i>B2</i>	numeric	Question answerer	<i>s</i> answered one question
<i>B3</i>	numeric	Star-crossed	<i>s</i> agreed or disagreed with a comment
<i>B4</i>	numeric	Comment	<i>s</i> wrote one comment
<i>B5</i>	numeric	Author-reply	<i>s</i> replied to a comment written about own question
<i>B6</i>	numeric	Follower	<i>s</i> followed one or more authors
<i>B18</i>	numeric	Leader	<i>s</i> had one or more followers
<i>B19</i>	numeric	Helper	<i>s</i> responded to one help request or more
<i>B23</i>	numeric	Verifier	<i>s</i> has confirmed one answer or more
BADGES THAT CAN BE EARNED MORE THAN ONCE			
<i>B7</i>	numeric	Good question author	per question authored rated as excellent five times or more
<i>B8</i>	numeric	Popular question author	per question authored that was answered ten times or more
<i>B9</i>	numeric	Discussed question author	per question authored that received two or more comments
<i>B10</i>	numeric	Commentator	<i>s</i> wrote five comments or more
<i>B11</i>	numeric	Critic	<i>s</i> agreed or disagreed with ten comments
<i>B12</i>	numeric	Rater	<i>s</i> submitted a rating for ten questions
<i>B13</i>	numeric	Scholar	<i>s</i> answered ten questions correctly
<i>B14</i>	numeric	Genius	<i>s</i> answered ten questions in a row correctly
<i>B15</i>	numeric	Einstein	<i>s</i> answered twenty questions in a row correctly
<i>B16</i>	numeric	Insight	<i>s</i> wrote two or more comments that are agreed with by someone
<i>B17</i>	numeric	Conversation	<i>s</i> replied to five comments about own questions
<i>B24</i>	numeric	Super scholar	<i>s</i> answered correctly a total of 50 questions
TIME SENSITIVE BADGES			
<i>B20</i>	numeric	I'll be back	<i>s</i> answered correctly ten or more questions, on each of three different days)
<i>B21</i>	numeric	Commitment	<i>s</i> answered correctly ten or more questions, on each of five consecutive days
<i>B22</i>	numeric	Obsessed	<i>s</i> answered correctly ten or more questions, on each of ten consecutive days
<i>B25</i>	numeric	Legend	<i>s</i> submitted a correct answer on 31 distinct days

TABLE 6.6: Features extracted from the Assessment data dataset (including the Wiki, where groups allocation was published)

Feature	Data type	Description	Example values
<i>Group</i>	enum	Group for coursework (as per the Student Wiki)	group_1...
<i>'Faculty Code'</i>	enum	Faculty where the student is registered	F7, F8
<i>Exam_Mark</i>	numeric	Exam mark (weighs 50% of the final mark)	0, ..., 100
<i>Assessment_Mark</i>	numeric	Average mark on coursework	0, ..., 100
<i>Final_Mark</i>	numeric	Final mark (late penalties included)	0, ..., 100



The next set of extracted features come from the assessment data files, as listed in Table 6.6. The additional datasets, generated outside PeerWise, which are used to augment the data with organisational<sup>8</sup> and performance information are shown in the schema of Figure 3.6.

### Engineered features

By contrast, the features listed in Table 6.7 were engineered rather than extracted. This has meant that some preprocessing was applied on to the file of schemas shown in Figures 3.5 and 3.6 to construct a vector associated to each student in the dataset. Most of the features engineered were inspired by the model, with a few exceptions. For example, *Exam\_Mark\_nominal* and *Final\_Mark\_nominal*, which are mere transformations of numeric data into nominal data (as per the degree-classifications rules at the University of Southampton). Another example are the features listed in the second section of the table (from *Followers* to *Answers\_given*), which were calculated through grouping by User, Follower or Author in the relevant files and counting for each distinct User\_ID.

More interesting were the sets of features inspired by the model as those required making some design decisions and more complex transformations to the data. For example, for *Questions\_made*, I made the decision to model them as communicative e-tivities, even though they primarily reflect interactions between learners and their learning material, through the production of self-authored MCQ. However, as mentioned earlier, since these MCQs are then offered to the peers so they can answer them and critique them, these are essentially communicative e-tivities, even if they never become part of a “conversation” (in which case they are *Lone\_questions*). When they do spark comments from others, then they become *Starting\_questions*. In order to establish which of the two kinds it is, the question ID is searched for in the Comments file, and if there is a match, it means that at least a comment was made on said question. A similar analysis can be made, guided by the model, to study what are the relationships to be found across the various files in the dataset schema and engineer such features. Therefore, the first set of questions in Table 6.7, ranging from *Questions\_made* to *Initiators\_Replies* are all communicative e-tivities.

There is an important comment to make with regards to non-communicative e-tivities, such as *Answers\_given*, which is an example of counts of learner engagement

---

<sup>8</sup>For example, the Groups\_Wiki information is as captured from the Student Wiki page for the module COMP 2213, which was used both years for students to self-organise themselves in groups. Available at: <https://secure.ecs.soton.ac.uk/student/wiki/w/COMP~2213-1516>, accessible though the departmental intranet.

TABLE 6.7: Features engineered from the PeerWise dataset for a given student  $s$  (with a unique  $User\_ID$  as per Table 6.5).

Feature	Data type	Description	Example/ Values
<i>Questions_made</i>	numeric	number of questions (MCQs) authored by $s$	0...20
<i>Comments_received</i>	numeric	number of comments received by $s$	0...21
<i>Starting_questions</i>	numeric	number of MCQs authored by $s$ that receive comments	0...12
<i>Lone_questions</i>	numeric	number of MCQs authored by $s$ that do not receive comments	0...8
<i>Comments_made</i>	numeric	number of comments made by $s$	0...25
<i>Replies_made</i>	numeric	number of replies made by $s$	0...14
<i>Initiators_Replies</i>	numeric	number of replies made by $s$ to comments on $s$ 's MCQs	0...12
<i>Followers</i>	numeric	number of students who follow $s$	0...3
<i>Following</i>	numeric	number of students followed by $s$	0...3
<i>Ratings_given</i>	numeric	number of times that questions $s$ have been rated for quality	0...82
<i>Avg_qual_ratings_given</i>	numeric	average quality rating given to questions by $s$	0...5
<i>Answers_given</i>	numeric	number of MCQs answered by $s$ (all attempts)	0...529
<i>0-Early_engagement_Question</i>	numeric	As per <i>Questions_made</i> but disaggregated by period	
<i>1-Easter_Question</i>	numeric		
<i>2-Exam_revision_Question</i>	numeric		
<i>3-Post_exam_Question</i>	numeric		
<i>0-Early_engagement_Answer</i>	numeric	As per <i>Answers_given</i> but disaggregated by period	
<i>1-Easter_Answer</i>	numeric		
<i>2-Exam_revision_Answer</i>	numeric		
<i>3-Post_exam_Answer</i>	numeric		
<i>0-Early_engagement_Comment</i>	numeric	As per <i>Comments_made</i> but disaggregated by period	
<i>1-Easter_Comment</i>	numeric		
<i>2-Exam_revision_Comment</i>	numeric		
<i>3-Post_exam_Comment</i>	numeric		
<i>0-Early_engagement_Ratings</i>	numeric	As per <i>Ratings_given</i> but disaggregated by period	
<i>1-Easter_Ratings</i>	numeric		
<i>2-Exam_revision_Ratings</i>	numeric		
<i>3-Post_exam_Ratings</i>	numeric		
<i>0-Early_engagement_Reply</i>	numeric	As per <i>Replies_made</i> but disaggregated by period	
<i>1-Easter_Reply</i>	numeric		
<i>2-Exam_revision_Reply</i>	numeric		
<i>3-Post_exam_Reply</i>	numeric		
<i>Exam_Mark_nominal</i>	enum	Classification from exam marks	first...
<i>Final_Mark_nominal</i>	enum	Classification from final marks	first...

with the material (i.e. answering MCQs). The dataset I received does not offer interval information, so they are all treated as if they were atomic. Therefore, there are no interval features in the PeerWise datasets used in this thesis.

There are, however, other features somewhat reflecting temporal information, which are those disaggregated by period. The periods are defined by the following milestones in the semester: that is, from coursework release to Easter, during Easter, after Easter but before the exam, and after the exam. These result in the features with the prefixes *0-Early\_engagement\_*, *1-Easter\_*, *2-Exam\_revision\_* and *3-Post\_exam\_*.

## 6.5 Clustering on PeerWise features

Having identified in the MOOC case what features to use (in Section 5.6), I use its equivalent feature set, the five dialogic features: starting posts, lone posts, first replies, initiator's replies and additional replies (i.e. SP, LP, FR, IR, AR, the communicative e-tivities in the model of learner engagement defined in Section 4.3).

There is a good correspondence (by design, after all, these features were engineered based on the model), despite a somewhat confusing overlap in nomenclature. All communicative e-tivities are comments in MOOCs, but in PeerWise there are questions, comments, and replies, each of these types being zero-, first-, and second-order replies as discussed in Section 6.3. After a visual inspection of Figures 4.6 and 6.2 it is evident that the correspondence amongst those listed in Table 6.7 is as follows:

- *Starting\_questions*: equivalent to **SP**
- *Lone\_questions*: equivalent to **LP**
- *Comments\_made*: equivalent to **FR**
- *Initiators\_Replies*: equivalent to **IR**
- *Replies\_made*: equivalent to **AR**

Therefore, to facilitate the comparison, from here onwards I will refer to the model-equivalence names rather than the name of the engineered features from the PeerWise dataset.

### 6.5.1 Size and coherence of resulting clusters

Similarly to how it was shown for MOOCs in Section 5.7.1, I investigate the size of the resulting clusters and their coherence. In order to do so, I first inspect the confusion matrices generated in the same way as described in Section 5.6.1, when the scikit-learn `DecisionTreeClassifier` was used to determine the goodness of the fit. Figures 6.3 and 6.4 show the confusion matrices for the classifier when predicting the seven clusters found by X-Means on the datasets related to the first and second cohorts of Interaction Design, respectively. These must be interpreted in the same way as described in Section 5.7.1.

On inspection, it is possible to make the following observations:

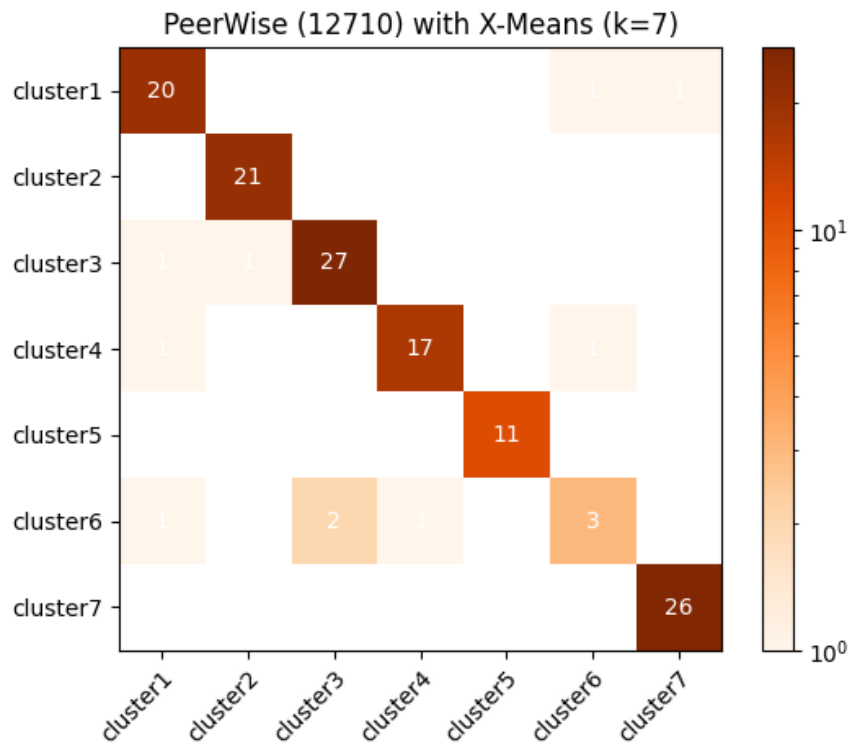


FIGURE 6.3: Confusion matrix plots for the scikit-learn DecisionTreeClassifier on the seven clusters found by X-Means applied to data from the first cohort using PeerWise (course 12710), with  $k = 7$

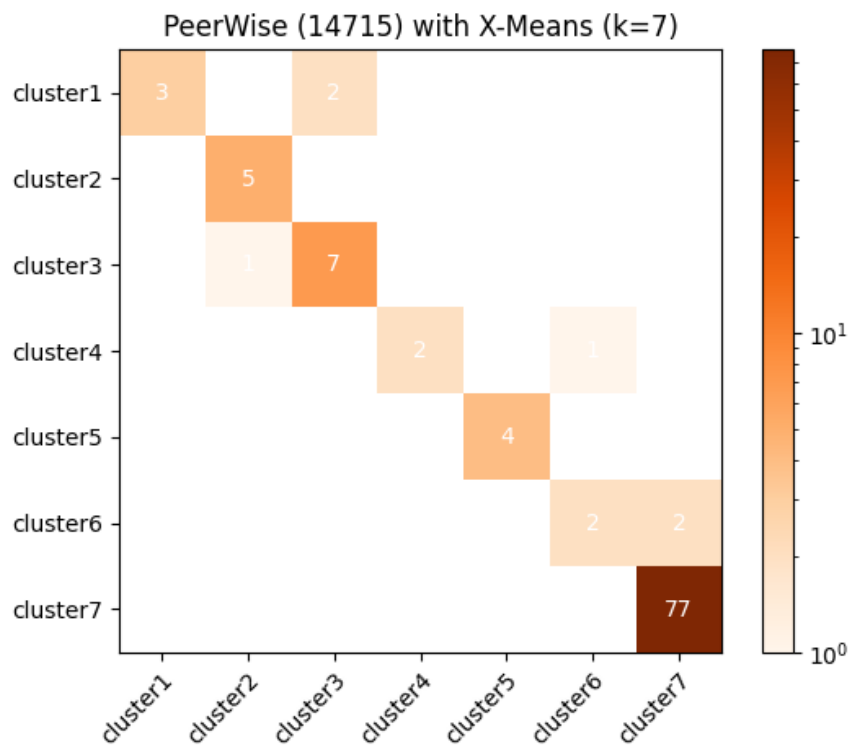


FIGURE 6.4: Confusion matrix plots for the the scikit-learn DecisionTreeClassifier on the seven clusters found by X-Means applied to data from the second cohort using PeerWise (course 14715), with  $k = 7$

### **With regards to size**

Unlike the confusion matrices for the MOOCs studied, these have two very different profiles. In the confusion matrix for the first cohort (Figure 6.3) the clusters found were fairly balanced, with an average of 16.3 learners and a standard deviation of 10, with only one cluster being uncharacteristic from the rest (cluster6) in that it had slightly fewer instances than the average minus the standard deviation. By contrast, the confusion matrix for the second cohort (Figure 6.4) was very imbalanced, as despite having a similar average (of 15 learners), the standard deviation was 28.35, largely due to the influence of only one cluster with a number of learners more than two standard deviations away from the mean (cluster7, with 77 learners, accounting for more than 73% of the whole dataset) similarly to what was observed in the MOOC case.

### **With regards to coherence**

The classification is very accurate in both cases, with diagonally-dominant confusion matrices with very little noise outside the diagonal indicating a handful of misclassifications. As in the MOOC case, the clusters are inherently coherent, as the high level of similarity between instances within the same cluster are highly-accurately predicted when subjected to the scikit learn `DecisionTreeClassifier`. This is evidently true for all clusters in both sets of results, though less so for the cluster labelled as cluster6 in both cases. Given this high inter-cluster coherence, the next logical step is to assign clusters meaningful names characterise each cluster in domain-interpretable terms.

## **6.5.2 Semantically chosen names for clusters in both courses**

Having established that the seven clusters found via unsupervised learning are highly coherent, I next inspect the resulting clusters to give them meaningful names based on the central measures of the instances within.

The box-and-whiskers plots shown in Figures 6.5 and 6.6 are interpreted in exactly the same way as explained in Section 6.5.2, with an important difference: the label for the y axis is not the number of “comments”, but of “communicative e-tivities”. In the case of MOOCs these terms could be used interchangeably because all of the communicative e-tivities in the FutureLearn platform are comments, irrespective of their place in the dialogue. However, for PeerWise, as we have seen in Section 6.3, the communicative e-tivities comprise questions, comments and replies instead.

### For the first cohort

The semantics of each cluster in this dataset are loosely based on the names of the groups by the heuristic in [Chua et al. \(2017\)](#), according to the mean and median values for the dialogic features as shown in Figure 6.5, just as in the MOOCs case. These names will allow for a comparison of engagement across platforms. The semantics for each cluster are as follows:

**cluster1** *Even more active social learners without turn-taking:* Similar to cluster2 and cluster4 below but with a slightly higher number of communicative e-tivities overall. These learners produce exactly four multiple-choice questions (MCQs), one of which is typically a starting post (SP) and three are lone posts (LP). This number of questions were the minimal engagement requirement as per the assessment design. Therefore, the only additional activity these learners engaged with, were replies (approximately five on average). All other dialogic features are zero (apart from the case of one outlier). This cluster has 22 instances.

**cluster2** *More active social learners without turn-taking:* As above but with fewer communicative e-tivities for the means and medians of all dialogic features. Both SP and LP are exactly two, meaning that, like in the cluster above, the learners here produced exactly the minimum requirement for MCQs and no more. There are 21 instances in this cluster.

**cluster3** *Loners:* All features are zero apart from LP for most of the instances (except for five outliers). The size of the cluster is 29 instances.

**cluster4** *Active social learners without turn-taking:* Learners in this cluster created exactly four MCQs, the minimum requirement. Three of these questions (SP) sparked comments from others, and one did not (LP). The median for all remaining dialogic features is zero, but not the means, as given the number of instances, they are sensitive to outliers. There are 19 instances in this cluster.

**cluster5** *Asocial learners:* All dialogic features are close to zero, but given the size of the cluster, the means are sensitive to outliers. There are 11 instances in this cluster.

**cluster6** *Active social learners:* In this group, both central measures for all dialogic features are greater than zero, except for initiators' replies (IR), for which the median is zero. There are 7 instances in this cluster.

**cluster7** *Initiators without replying.* This group has non-zero starting posts (SP) yet a zero median for all kinds of replies (FR, IR and AR). The mean is very close to zero, but non-zero because of the outliers. There are 26 instances in this cluster.

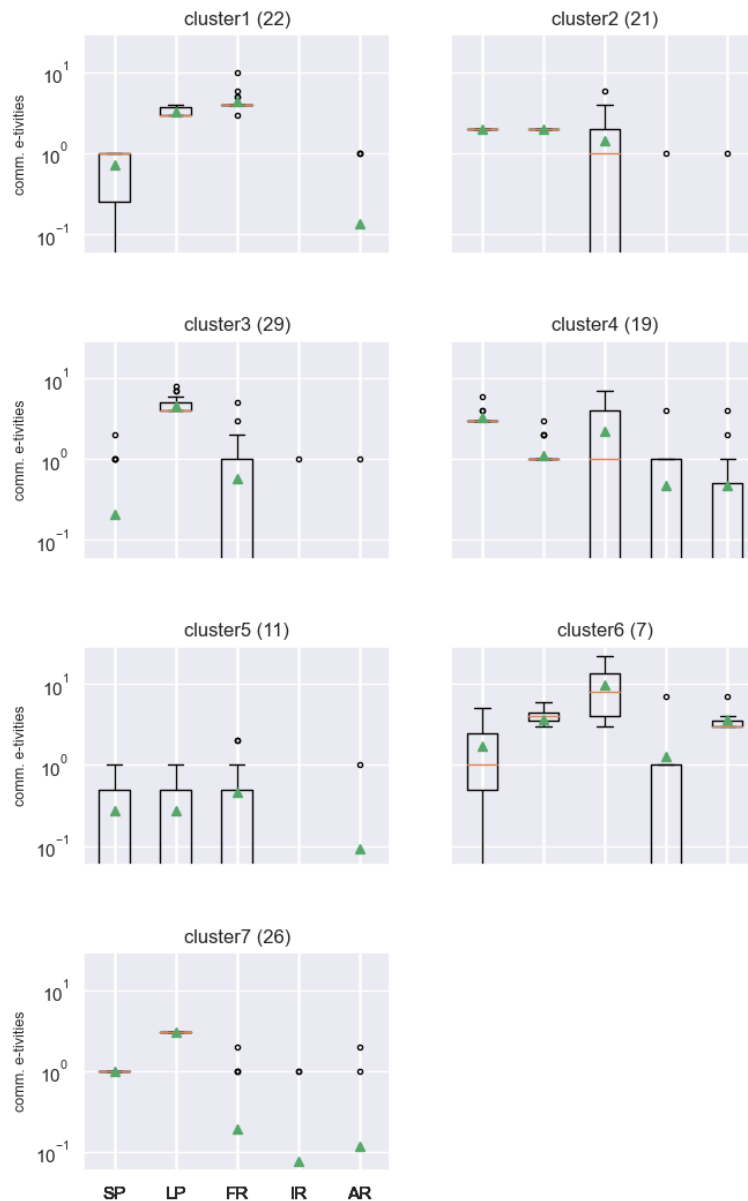


FIGURE 6.5: Box-and-whisker plots for the dialogic features on clusters found by the X-Means clustering algorithm on data from the first cohort using PeerWise (course 12710), with  $k = 7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1** : Even more active social learners without turn-taking; **cluster2** : More active social learners without turn-taking; **cluster3** : Loners; **cluster4** : Active social learners without turn-taking; **cluster5** : Asocial learners; **cluster6** : Active social learners; **cluster7** : Initiators without replying.

### For the second cohort

For the second cohort, I analysed the central measures of the clusters shown in Figure 6.6, guided by the categories in Section 5.3 and similarly applied above. One small variation with respect to the box-and-whiskers plots shown so far is that the scale used is linear rather than logarithmic. In this context it is less appropriate to use the logarithmic scale that was used in the MOOC context (where on occasion a handful of posts had to be shown against over a hundred first replies in the same plot.) However, more importantly, because for this dataset there are so many features with values equal to zero, the plots become more difficult to interpret as the scale does not include zero by definition, but starts with a very small number instead.

This resulted in the following cluster names:

**cluster1** *Active social learners*: In this group, the central measures most dialogic features (except LP) are greater than zero, but with a lower level of activity than that shown in cluster3 below. There are five instances in this cluster.

**cluster2** *Initiators without replies*: All dialogic features are zero, apart from starting posts (SP), meaning that these learners produced MCQs that attracted comments, but they never replied to such comments. There are five instances in this cluster.

**cluster3** *More active social learners*: Similar to cluster1, only with a higher level of comment activity. Particularly striking are the central measures (median, eight first replies, and mean, nine) and the spread, with up to 25 first replies in the 75<sup>th</sup> centile. There are eight instances in this cluster.

**cluster4** *More active repliers*: Similar to cluster5 below. All dialogic features are zero, apart from first replies (FR) and additional replies (AR), meaning that these learners left comments on MCQs that someone else produced, but not created any themselves. There are three instances in this cluster.

**cluster5** *Repliers*: All dialogic features are zero, apart from first replies (FR). More specifically, that these learners left one comments on someone's MCQ and never engaged in dialogue again within this course. There are four instances in this cluster.

**cluster6** *Loners*: SP and IR are zero and all other dialogic features are non-zero (albeit with very low values). There are four instances in this cluster.

**cluster7** *Asocial learners*: All dialogic features are exactly zero for all learners in this cluster, which also is the biggest of the whole dataset, 77 instances.



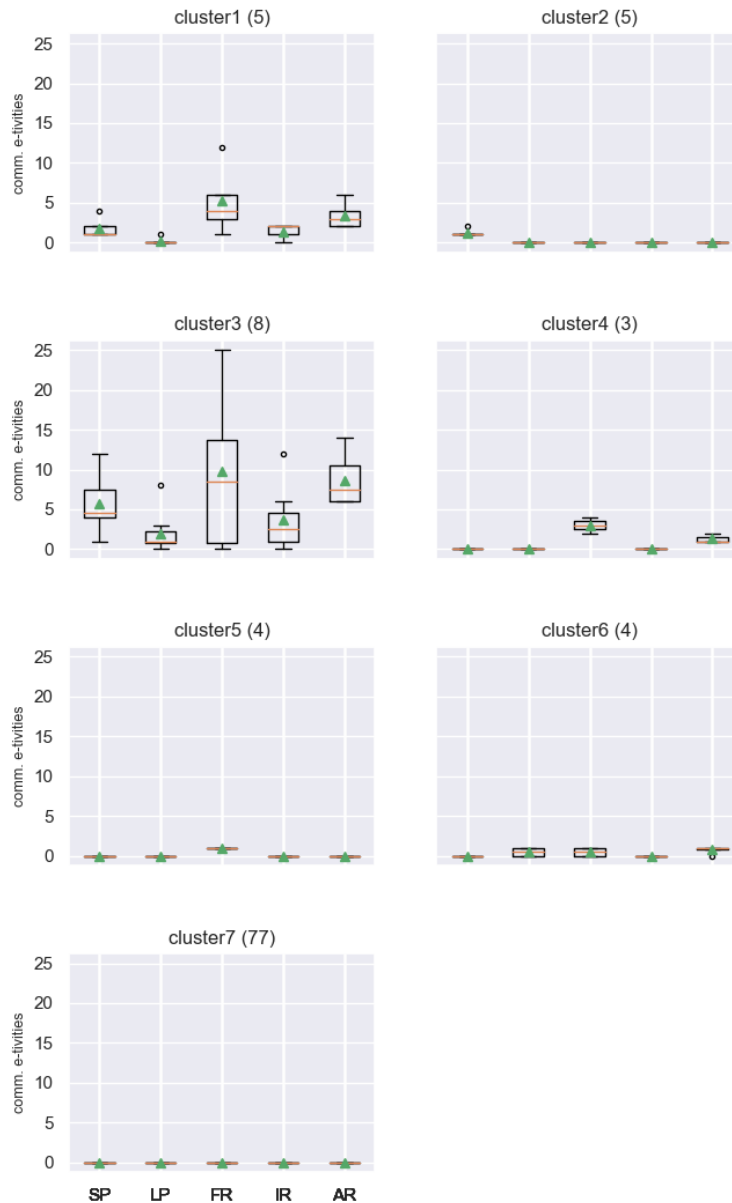


FIGURE 6.6: Box-and-whisker plots for the dialogic features on clusters found by the X-Means clustering algorithm on data from the second cohort using PeerWise (14715), with  $k = 7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1** : Active social learners; **cluster2** : Initiators without replies; **cluster3** : More active social learners; **cluster4** : More active repliers; **cluster5** : Repliers; **cluster6** : Loners; **cluster7** : Asocial learners.

Given that there seem to be several clusters with similar semantics, and four very distinct ones, for these datasets it makes sense to show the box-and-whiskers plots produced when the X-Means clustering algorithm is applied but reducing the number of clusters to four. Figures 6.7 and 6.8 show the resulting box-and-whisker plots. In this case, the difference in observable behaviour between classes is even more noticeable: though the second cohort does have many asocial learners (83) and the first one has none, only “loners” (54). These students were engaging in PeerWise under the behavioural constraints imposed by the assessment. Learners in the first cohort were nudged to create four questions as marks were awarded for doing so (see Appendix F), so there is a class of minimal engagement whereas for the second cohort it was an entirely voluntary activity, and hence they behaved very much like the MOOC learners. Though fewer, the active social learners in the second cohort created many more posts than those in the first cohort and also engaged in replies much more.

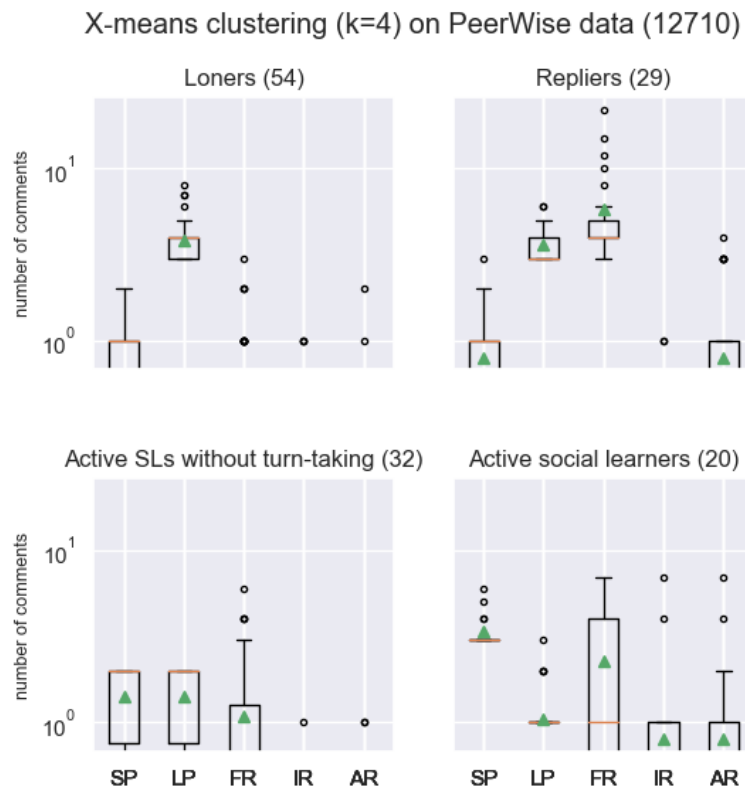


FIGURE 6.7: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the PeerWise course 12710, with k=4

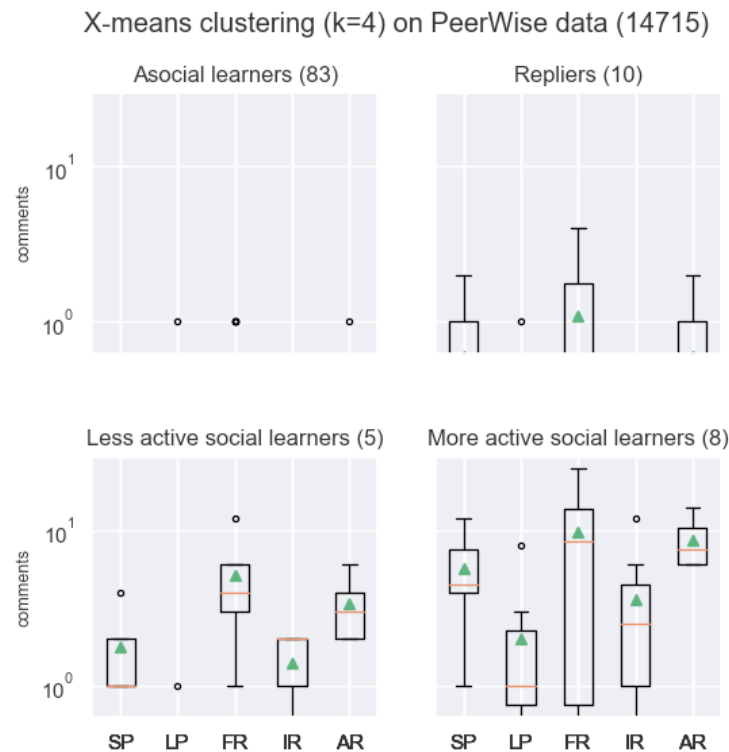


FIGURE 6.8: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the PeerWise course 14715, with  $k=4$

## 6.6 Reflecting back to MOOC analysis

Having chosen meaningful cluster names to learners in both cohorts of Interaction Design based on their engagement in PeerWise, I can now make a cross-platform comparison of said engagement. This is done in Tables 6.8 and 6.9. Here I display the absolute numbers of learners in both MOOCs (when learners in all runs were aggregated, as shown in Figures 5.11 and 5.12), together with those in both cohorts of Interaction Design using PeerWise (in Figures 6.5 and 6.6), according to each of the semantic analyses of central measures of clusters found by X-Means with  $k=7$  in a like-for-like comparison (despite results with  $k=4$  having a slightly better fit, as shown through the information retrieval metrics for the classifier, detailed in Appendix J).

In the previous sections I made a couple of important observations. Firstly, the difference between the behaviours of the first and the second cohort of Interaction Design, and secondly, the similarity between the second cohort behaviours and those observed in MOOCs. In this ex post facto quasi-experiment, the independent variable, that is presumed to be responsible for these observation, is whether or not there were incentives for participation in PeerWise.

TABLE 6.8: Comparison of numbers of learners in each of the semantic classes for the clusters found by X-Means in both peer-supported digital environments: For Portus (all runs), Understanding Languages (all runs), and both cohorts of Interaction Design using PeerWise (courses 12710 and 14715 respectively).

Classes found	Portus	UL	PeerWise	
	(all)	(all)	12710	14715
1- asocial learners	14,602	93,851	11	77
2- loners		10,827	29	4
3- initiators without replying	1,108	7,268	26	5
4- initiators who respond	25			
5- replier				4
5a- more active replier				3
7- ASL without turn-taking		4,843	19	
7a- more active SL without turn-taking			21	
7aa- even more active SL without turn-taking			22	
8- active social learners	150	691	7	5
8a- more active social learners	32			8
8b- ASL who do not give additional replies	313	2,679		
8bb- more active SL who do not give additional replies	114	1,313		

TABLE 6.9: Percentages of learners in each of the semantic classes for the clusters found by X-Means in both peer-supported digital environments. Due to rounding, percentages may not add up to 100%.

Classes found	Portus	UL	PeerWise	
	(all)	(all)	12710	14715
1- asocial learners	89%	77%	8%	73%
2- loners		9%	21%	4%
3- initiators without replying	7%	6%	19%	5%
4- initiators who respond	0%			
5- replier				4%
5a- more active replier				3%
7- ASL without turn-taking		4%	14%	
7a- more active SL without turn-taking			16%	
7aa- even more active SL without turn-taking			16%	
8- active social learners	1%	1%	5%	5%
8a- more active social learners	0%			8%
8b- ASL who do not give additional replies	2%	2%		
8bb- more active SL who do not give additional replies	1%	1%		

Having incentives directly linked to the assessment (however small, such as the 5% allocated for it in the first cohort), together with the nudges created by group membership, is seen to have caused an effect. This is particularly true for features SP and LP (starting posts and lone post, or more precisely, *Starting\_questions* and *Lone\_questions* as engineered from PeerWise), as these were non-zero for the vast majority of learners in the cohort. The fact that the sum of these two features was exactly four for the vast majority of learners, as shown in Figure 6.5, suggests that these learners interacted through the platform only as required by the constraints imposed by the assessment design.

Freed from these behavioural constraints, however, the second cohort behaved more like online learners, in that the vast majority did not create any MCQs (hence SP and LP were both zero for them) and only a dedicated few created as many as 25 MCQs, as well as fully engaging with the conversational capabilities in the platform, by way of making comments and replies, including initiators' replies. This is what I call the Theory of Behavioural Constraints:

*In a peer-supported digital environment complementing face-to-face instruction, learners will exhibit behaviours typical of online learners unless their behaviour is constrained by effective interventions.*

From this theory it follows that some interventions (such as incentives within assessment, or exposure to social salience) can be instrumental for moving learners from low-activity classes to higher-activity ones. When these are incorporated into the learning design, a significant behaviour change can be effected by even simple nudges (e.g. of persuasion, provision of information, or use of social norms and salience, as described in Section 2.1.1). In fact, the learning design in FutureLearn incorporates, as seen in Section 2.2.1, a variety of nudges based on an explicit pedagogy of conversational learning. I believe these nudges to be directly responsible for the higher level of observed learner engagement than in some other x-MOOCs that are not based on a conversational framework, as noted by [Sharples and Ferguson \(2019\)](#).

However it must be noted that, in practice, more powerful interventions are much more challenging, if not impossible to implement, in a fully-online learning environment such as MOOCs. The face-to-face context itself further facilitates the effectiveness of behavioural constraints, in terms of social salience for example, in the form of peer-pressure from groups formed in-person. However, there are other factors at play, such as the intrinsic motivation from the student to do well in a course given the much larger stakes, as the consequences of ultimately failing a course are loss of a huge financial investment (particularly in fee-paying contexts, such as in British universities) but also in time, effort and even emotionally. The stakes are much lower in online courses like MOOCs, which may not only be free and shorter, but offered several times a year, and therefore, offering many opportunities to fail and start again. Hence some serious interventions in the form of disincentives or incentives (in the harsher end of the spectrum in Table 2.1) may be difficult to implement in MOOCs.

My research found evidence that learners in a face-to-face course exhibit behaviours within peer-supported digital environments that are essentially the same as those in online courses, though they may exhibit a different profile due to behavioural constraints

being in place. The second cohort of Interaction Design showed that when these constraints were removed, the behaviour profile became closer to that in the MOOCs I studied, with a large majority of asocial learners and much fewer active social learners.

On this point, however, it is worth commenting on the number of active social learners in the second cohort of students. At 13%, these constitute a larger proportion than the 90-9-1 rule would predict. I posit that this is explained by not having removed all of the behavioural constraints at play. As mentioned above, there are other factors at play, which are not directly observed, such as peer-pressure, sense of belonging and financial incentives, as indeed, the internal motivation of the high-achieving learners in this group, which drove them in the first place to embark on these studies at this university.

## 6.7 Summary and conclusion for this chapter

This chapter presented results on an ex-post facto quasi-experiment of students using PeerWise as the peer-supported digital environment within the Computer Science module *Interaction Design*. The study comprised two consecutive cohorts, in 2015/16 and 2016/17, subjected to different assessment conditions, by which learners in the first group were incentivised to author multiple-choice questions in PeerWise, whereas those in the second group were not.

I then used the model of learner engagement defined in Chapter 4, to explain an example of a synthetically-created dataset and applied a feature engineering process similar to the one presented in Chapter 5 but based on the information retrievable from the PeerWise dataset. In doing so, I found answers to the research question **RQ3** (*What does a data-driven approach to learner interactions reveal about learning engagement within the PeerWise digital environment for face-to-face instruction?*), as follows:

The clusters of learner engagement identified are shown in Table 6.10. These are a subset of those identified in the MOOCs which were shown in Table 5.16. This table highlights the differences in behaviour profiles between two cohorts, where the first one was incentivised to participate in PeerWise via rewards incorporated in the assessment design. Whilst the first cohort saw a larger overall uptake of the conversational affordances of the software (with a comparatively low number of asocial learners), these interactions tended to be shallower in terms of number of turns taken in the conversation, than in those the one freed from the behavioural constraint to participate: there was a larger number of initiators without replying (18% versus 5%), and a total of active social learners without turn-taking of 44% with three different levels of intensity, which were not observed in the cohort that was free from assessment incentives. This

suggests that the intervention was at least partly responsible for a redistribution of the naturally-inclined to be asocial learners into varying levels of interaction observed.

TABLE 6.10: Comparative table of clusters found in PeerWise data with X-Means ( $k = 7$ ) in each cohort of Interaction Design (PeerWise courses 12710 and 14715).

Identified categories of learner engagement	First cohort	Second cohort
1-asocial learners	8%	73%
2-loners	21%	4%
3-initiators without replying	18%	5%
5-replier	4%	4%
5a-more active replier		3%
7-ASL without turn-taking	13%	
7a-more active social learners without turn-taking	15%	
7aa-even more active social learners without turn-taking	16%	
8-active social learners	5%	5%
8a-more active social learners		8%

Finally, this chapter also provided answers to the research question **RQ4** (*Is learner engagement different in different kinds of peer-supported digital environments, be it a complement to face-to-face instruction, or a fully online course?*), through the comparison of the learner engagement in both environments using the same model-based, engineered features. Given that the cohort that was free of the behavioural constraint to interact through the peer-learning environment did so in a similar way to that observed in MOOCs (c.f. Table 6.9), I formulated the theory of behavioural constraints by which “In a peer-supported digital environment complementing face-to-face instruction, learners will exhibit behaviours typical of online learners unless their behaviour is constrained by effective interventions.”





## Conclusion

*No te dejes confundir  
Busca el fondo y su razón  
Recuerda, se ven las caras  
Pero nunca el corazón.*

Rubén Blades, “PLÁSTICO”,  
In *Siembra*, Fania Records, 1978.  
In the Latin Grammy Hall of Fame since 2007.

A song from my childhood includes the lyrics<sup>1</sup> in the epigraph. These words, deemed as received wisdom amongst many of my Latin American contemporaries, also apply to the analysis of digital traces of learning activity: the reasons behind specific behaviours are not directly observable, only their external manifestation. Hence, though we cannot truly classify learners, we can do so with their engagement in our courses according to the digital traces of their learning activity. In Chapter 4, I compared these digital traces to silhouettes of dancers behind a screen (Figure 4.2). The model of learner engagement defined in that chapter was used to examine two different environments (with different “dancers”, performing behind different kinds of screens and objects). In doing so, it became possible to use a common language to discuss underlying reasons for the differences in observed behaviours, between iterations of the same course, and between courses, irrespective of the platform used to capture traces of their learning activity.

This chapter is organised as follows: Section 7.1 presents the summary of this thesis. Section 7.2 returns to the research questions posed in Chapter 1 and explains how

---

<sup>1</sup>“Don’t let yourself get confused / search for depth and for its reason / remember, you see the faces / but you never see the heart” (translated from Spanish).

this thesis provided answers for them. Section 7.3 outlines the contributions as per the research framework followed in this thesis. Section 7.4, offers a discussion on the limitations of this research, and some avenues for future work are discussed in Section 7.5. Finally, this chapter ends with some concluding remarks, offered in Section 7.6.

## 7.1 Summary of this thesis

Chapter 1 outlines the motivation for this thesis and the purpose of this research, as articulated in the four research questions which I revisit at the end of this summary. In Chapter 2, I considered the fundamentals to learning and engagement, including brief descriptions of philosophies of learning, teaching and peer learning as we move towards a learner-centered paradigm, catalysed by the adoption of digital technologies in education. This led into the challenges of doing so at scale and how socio-constructivist MOOCs (and in particular FutureLearn MOOCs) are exponents par excellence of the social and collaborative aspects of learning, and the importance of a learner-centred approach. Selected literature on MOOC research was reviewed for approaches to categorise learners and it was presented comparatively in Table 2.3. This was similarly done on PeerWise research and Table 2.5 presents reported uses of PeerWise and features engineered from this data. Works in other peer-supported learning environments were also reviewed. A whistle-stop tour on learning analytics, feature engineering and clustering was given, as well as on the fundamentals of interval algebra, as these were all knowledge upon which I built work presented in later chapters.

In Chapter 3, I detailed the methodology used in this thesis, following a data science approach. but also to inform a quantitative, data-driven approach, and shed light on the semantic interpretation of its results. Steps in a data science pipeline were used as a high-level explanation of this research process, with particular attention to data collection and cleaning, although specifics on feature engineering and deployment were appropriately deferred to subsequent chapters.

In Chapter 4, I defined the model of learner engagement in peer-supported digital environment, using interaction to operationalise engagement, in particular considering communicative and non-communicative activities. In this thesis, this model was instrumental to overcome the differences in data representation and organisation of the datasets associated to the environments under study. The mechanism by which it did so was by informing the feature engineering process and the analysis of the findings.

In Chapter 5, experiments on datasets from MOOCs from the University of Southampton were conducted, and the whole learning analytics process was detailed, starting

from feature engineering and reduction of dimensionality, followed by the selection of both a good clustering algorithm and a suitable number of clusters to separate learners into. The clusterer's fitness was assessed using a supervised learning approach (classification) with the resulting clusters set as ground truth classes. The classification outputs were reported via confusion matrices, and the resulting clusters were inspected using box-and-whisker plots through which central measures and dispersion were assessed before being renamed into semantic classes. These newly-found semantic classes were compared to those suggested by a heuristic-based method (Chua et al., 2017), allowing for the identification of four main categories of learners based on their engagement. Having done the complete analysis at a individual-runs-of-a-MOOC level also allowed for the identification of inter-run variations of learner behaviours that were possibly induced by changes in the learning design of the MOOCs and affordances of the platform, such as email notifications.

Chapter 6 presented the application of the same processes detailed in the previous chapter, but using PeerWise data. A comparison of the findings from the data analysis in two consecutive cohorts of a face-to-face instruction course was possible.

## 7.2 Answering the research questions

In this Chapter I summarise how this research provided answers to the following four research questions:

**RQ1** *How can learner engagement be meaningfully compared across peer-supported digital environments?*

The model of learner engagement formulated in Chapter 4 provides an answer to this question as it allows the comparison of behaviours as manifested in each of the environments. As per the conclusions in Chapter 4, this model provides a common language to express relationships captured in the logged interactions within diverse peer-supported digital environments.

More specifically, it accomplishes that by considering learning activities in peer-supported digital environments as being either communicative or non-communicative. Each electronically-captured activity (*e-tivity*) can be described through a triple  $\langle a, l, t \rangle$ , to indicate an activity  $a$  ('what?') performed by a learner  $l$  ('who?') at a time  $t$  (either a timestamp or a period, 'when?').

The model considers five types of communicative activities, each of which fall into one of three types according to their place in a conversation, namely: zero-order replies comprising starting posts (SP) and lone posts (LP); first-order replies, comprising first replies (FR), and some LP (when they are ‘replies to self’ if no others are in the conversation); and second-order replies, comprising initiators’ replies (IR), further replies (FR) and some LP as before.

In the model, communicative activities are considered to take place over a period of time and are therefore governed by a variation of Allen’s algebra of intervals based on their timestamp, with the variation allowing for the inclusion of ongoing or abandoned activities, for which there is a known starting time but no end time.

The above abstractions for e-tivities allow navigating representational challenges across heterogeneous educational datasets. For example, in FutureLearn MOOCs all communicative e-tivities are logged in a `comments.csv` file, whereas in PeerWise there are three separate files, `questions.csv`, `comments.csv` and `replies.csv`. The same types of activities can be found in both environments, though. In PeerWise a ‘question’ is either an SP or an LP, a ‘comment’ is a FR or an LP, and a ‘reply’ an IR or an AR; whereas in FutureLearn they are all ‘comments’. However, they are just different expressions of the same conversational elements, and the model provides a framework to disentangle them. Further, the model can be used to inform feature engineering to increase interpretability in data-driven analyses on such heterogeneous data.

**RQ2** *What does a data-driven approach to learner interactions reveal about learning engagement within FutureLearn MOOCs?*

The application of unsupervised learning algorithms such as clustering (with X-Means providing a significantly superior performance amongst those I experimented with) allows the discrimination of classes of learner interaction behaviours naturally occurring in FutureLearn MOOCs. Through this method I was also able to identify between-runs cohort-wide variations in behaviour such as the disappearance (or significant reduction) of somewhat passive behaviour classes that gave way to the emergence of higher-activity classes; this particular change coincided with the incorporation of a new platform affordance by which learners received email notifications when others commented upon their posts. These simple nudges seem to have caused behaviour change in learners. Finally, a data-driven approach is superior to heuristic-driven approaches in capturing nuanced behaviours defined by the intensity of interactions, though it was able to do so because of the robustness of the features based on the model of learner interactions which in turn was heavily informed by the heuristics by [Chua et al. \(2017\)](#).

The semantics for the clusters in these plots are from the following classes of learner, as interpreted according to their median activity levels for starting posts (SP), lone posts (LP), first replies (FR), initiators replies (IR) and additional replies (AR). On inspection, clusters in these runs were found to be in as many as sixteen different categories, which extend those identified by [Chua et al. \(2017\)](#), as shown in [Table 7.1](#).

TABLE 7.1: Summary comparative table of clusters.

Categories found in most courses	Clusters found with X-Means ( $k = 7$ ) in at least one run of Portus or Understanding Language	Social learner groups in the heuristic by <a href="#">Chua et al. (2017)</a>
Asocial learners	1-asocial learners	N/A
'Loners'	2-loners	Loners
	2a-more active loners	
(Active) social learners without turn-taking	3-initiators without replying	Initiators without replying
	3a-more active initiators without replying	Initiators who respond
	4-initiators who respond	Repliers
	5-replier	Reluctant active social learners
	6-reluctant ASL	Active social learners without turn-taking
	7-ASL without turn-taking	
	7a-more active SL without turn-taking	
Active social learners	7aa-even more active SL without turn-taking	Active social learners
	8-active social learners	
	8a-more active social learners	
	8aa-even more active social learners	
	8b-ASL who do not give additional replies	
	8bb-more active SL who do not give additional replies	

**RQ3** *What does a data-driven approach to learner interactions reveal about learning engagement within the PeerWise digital environment for face-to-face instruction?*

I applied a data-driven approach to study learner engagement of students in two consecutive cohorts of a course in a face-to-face instruction context. This revealed the effect of an intervention applied to one of these two as an opportunistic, ex post facto quasi-experiment. Thanks to a domain-informed, model-based feature engineering process, the X-means clustering algorithm identified several distinct behaviours amongst these learners, which are listed in [Table 7.2](#).

TABLE 7.2: A summary table of comparisons between clusters found amongst Interaction Design students (more details in Table 6.10).

Identified categories of learner engagement	First cohort	Second cohort
1-asocial learners	8%	73%
2-loners	21%	4%
3-initiators without replying	18%	5%
5-replier	4%	4%
5a-more active replier		3%
7-ASL without turn-taking	44%	
8-active social learners	5%	13%

**RQ4** *Is learner engagement different in different kinds of peer-supported digital environments, be it a complement to face-to-face instruction, or a fully online course?*

A data-driven approach to learner interactions, informed by the above-defined model was able to identify across all runs of the MOOCs studied as well as within both iterations of the face-to-face courses:

- A cluster of *asocial* learners, who did not do any posts (or questions, in the PeerWise case), comments or replies. In relation to the model of learner interactions, the trees associated to each of these learners would be “stumps” (or nodes without any children), i.e. have depth zero. These learners would not appear in the trees associated to other learners either, as they did not take part in any conversation. For the majority of the cases, this was the dominating behaviour, and it was observed in MOOCs as well as in PeerWise (when participation was optional rather than compulsory).
- A cluster of *loners*, who despite having made contributions (posts, in the case of MOOCs, or questions, in the case of PeerWise), these tended not to spark any comments from peers. Therefore, learners in this cluster did not engage in conversations within the environment. In relation to the model of learner interactions, the trees associated to each of these learners would tend to have depth one, containing all the lone posts they made. These posts are called zero-order replies in the model.
- A cluster of less-engaged learners. This manifested in MOOCs and PeerWise in slightly different behaviours, however. For MOOCs they were categorised as *active social learners without turn-taking* and included those who tended to initiate conversations (some, if not all, of their posts were commented upon), and also

replied to others. However they tended not to reply to comments received about their own initiating posts, and therefore the associated trees of interactions would tend to have depth two, with the first level containing all their initiating posts and lone posts (or questions), and the second containing the comments received on them (first-order replies). In addition, these learners also feature in some of their peers' trees, as they would have made comments to their posts or questions. This cluster was replaced by a cluster of even less active learners, who either initiated or replied in PeerWise, but not both.

- A cluster of *active social learners*, comprising those who, in addition to engaging in the behaviours by the other groups, would also reply under their own initiating posts (do additional replies, or second-order replies as per the model). The associated interaction trees would have nodes at every level (and therefore have depth three).

In addition to the above, other clusters of nuanced behaviours were observed in MOOC data which were not observable in PeerWise, presumably due to scale, as PeerWise learners were 150 in average, whereas in the MOOC studied there were thousands or tens of thousands. Amongst these large datasets a handful of individuals exhibited behaviours which even though they were identified by the clustering algorithm as distinct enough to form a category in their own right (rather than being an outlier to other categories, say), statistically speaking, these were rare and therefore unlikely to be observed in a much smaller sample as indeed the face-to-face student cohorts were. Table 6.9 presented the full list of behaviours found by the clusterer in each of these environments and the proportions of learner falling within each category. This is reproduced here for convenience as Table 7.3:

The examination of the main classes listed above led to interesting insights in both contexts, in particular when considering consecutive iterations of the same course. Interventions (in either the learning design, assessment, or even in the platform's affordances) do have an effect on the overall behaviour of the learners.

Further, this thesis found that, left to their own devices, i.e. without incentives for participation being an explicit element in the assessment of the course, learners' engagement in peer-supported digital environments complementing face-to-face courses tend to be like that for online learners. Conversely, this means that interventions (by way of offering incentives for participation) are effective in migrating learners who would otherwise behave asocially. In my quasi-experiment, the *asocial learners* category was displaced when the participation in PeerWise was compulsory, and the categories of *loners* and *active social learners without turn-taking* emerged instead. I called this a theory

TABLE 7.3: Percentages of learners in each of the semantic classes for the clusters found by X-Means in both peer-supported digital environments. Due to rounding, percentages may not add up to 100%.

Classes found	Portus (all)	UL (all)	PeerWise	
			12710	14715
1- asocial learners	89%	77%	8%	73%
2- loners		9%	21%	4%
3- initiators without replying	7%	6%	19%	5%
4- initiators who respond	0%			
5- replier				4%
5a- more active replier				3%
7- ASL without turn-taking		4%	14%	
7a- more active SL without turn-taking				16%
7aa- even more active SL without turn-taking				16%
8- active social learners	1%	1%	5%	5%
8a- more active social learners	0%			8%
8b- ASL who do not give additional replies	2%	2%		
8bb- more active SL who do not give additional replies	1%	1%		

of behavioural constraint, which posits that face-to-face learners are not fundamentally different to online learners, yet they do exhibit different behaviours when their context is more rigorously managed through interventions.

## 7.3 Contributions of this thesis

This thesis followed the research framework presented as Figure 1.2. For ease of reference against the list that follows, this is reproduced conveniently as Figure 7.1, with an adaptation, to indicate the Sections related to each of the outputs from the processes applied throughout this research. Sections 7.3.1 to 7.3.5 highlight the main contributions of this thesis, depicted in this Figure in white boxes.

### 7.3.1 A platform-agnostic model for analysis of learner engagement

The model defined in Chapter 4 supports the analysis of complex, heterogeneous data in a simplified way, as in its abstraction it considers the fine-granularity activities that can be captured in a peer-supported digital environment. To the best of my knowledge, it is the first of its kind, in terms of mathematical rigour, simplicity, expressiveness, and completeness (despite its limitations) as it incorporates both communicative and non-communicative electronically-captured activities. This a valuable contribution, since much learning analytics research rest on the assumption that digital traces of activities



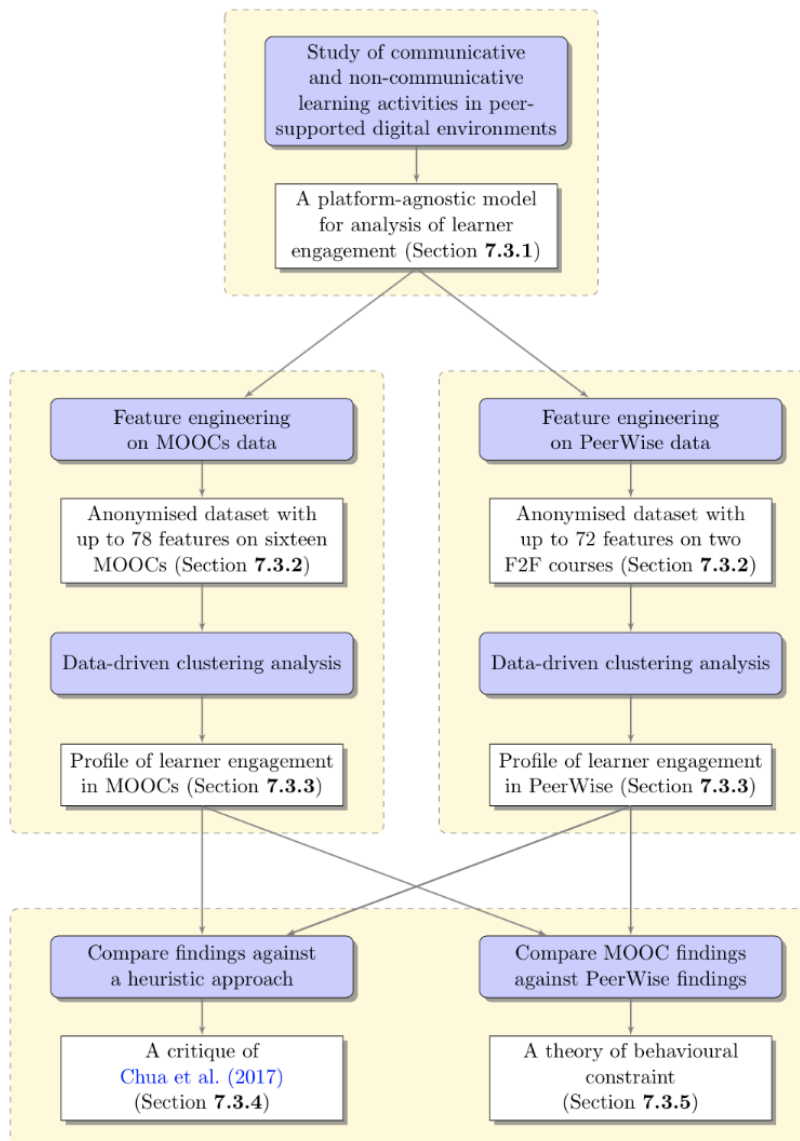


FIGURE 7.1: The research framework diagram from Figure 1.2, redrawn for convenience. The main contributions of this thesis are listed in white boxes, and the sections numbers listed in bold).

are useful proxies for behaviours, yet there is a gap in supporting this assumption with theoretical models of digital traces of learner activity. Further, this model is an important contribution in practical terms too, as it has proven its ability to inform the feature engineering process.

### 7.3.2 Anonymised datasets with up to 78 features on sixteen MOOCs and 72 features on two face-to-face courses

I generated these interim outputs as part of my data-driven approach to discover profiles of learner engagements on comparable datasets (i.e. datasets with the same feature sets). However, by uploading them into an open repository, these can also be used by other researchers wishing to reproduce my methods or answer other research questions that are suited to this kind of data. This a valuable contribution given the current lack of open, suitably anonymised data for learning-analytics research which is still a significant challenge in the field.

### 7.3.3 Profiles of learner engagement in MOOCs and PeerWise

I applied a data-driven approach, using unsupervised learning algorithms on the datasets previously generated. Unsupervised learning approaches, such as clustering, has been used successfully in the literature to identify underlying classes of learner behaviours both in FutureLearn MOOCs (Ferguson and Clow, 2015b) and elsewhere (Kizilcec et al., 2013). My work adds to this literature, and differentiates itself from previous approaches in that a theoretical model of learner engagement in peer-supported digital environments was used to guide the feature-engineering process prior to the clustering.

More specifically, I selected and used the X-Means clustering algorithm to determine seven clusters for each iteration of the courses under study (both MOOCs and in F2F instruction), obtaining as a result interpretable profiles of learner engagement for each environment. Further, it has evidenced the ability to identified nuanced behaviours which can be missed in heuristic-based approaches like that by Chua et al. (2017) and others as detailed below.

### 7.3.4 A critique of Chua et al. (2017)

The dialogic categorisation of learners by Chua et al. (2017) is a good heuristic that takes into account the role of learners in the conversations, which is particularly useful in FutureLearn MOOCs, where both their research and much of mine are situated, but has wider applicability. I was able to apply the heuristic (as described by these researchers) in sixteen courses, and was able to identify the same groups as they did, following a not-too-similar distribution of membership of learners to each heuristic class. However, in applying a data-driven approach, via unsupervised learning, it became evident that

their heuristics were too tight to explain all the behaviour in the MOOCs I studied. In particular, the heuristics do not capture the nuances described in Section 7.3.3, and in many cases, for the clustering algorithm it was more important the level of activity learners engaged into, than whether they had posted a certain type of post.

### **7.3.5 A theory of behavioural constraint**

The results from applying the processes of feature extraction and clustering analysis to data from two very different peer-supported digital environments (FutureLearn MOOCs and PeerWise) suggests that when incentives for participation are removed, it is possible to observe similar behaviours. More precisely, I observed that in such situations, learners in F2F instruction interact in peer-supporting learning environments in a very similar way to those in MOOCs.

Incentivising participation (as done for the first cohort of students using PeerWise in COMP 2213) led the majority of students to exhibit question-contribution behaviour above a minimal engagement threshold. Yet when this requirement was removed, a large majority of students did not engage in neither question creation nor comments, with only a handful of learners evidencing very high levels of engagement, in a manner approximately consistent to the engagement predicted by the 90-9-1 rule and also seen in online learners. However, the number of active social learners was slightly larger than this rule would predict as there are other behavioural constraints in place by virtue of the students being in a face-to-face environment and subjected to peer-pressure and other non-removable constraints.

### **7.3.6 Publications and talks**

In addition to the main contributions of this thesis, discussed above, other contributions include my communications to the academic community on various aspects of this research, some of which are listed in my declaration of authorship, in pagexx. The following list includes non peer-reviewed talks, organised by their themes in relevance to this thesis.

#### **On learning in MOOCs compared with face-to-face instruction**

In these presentations I talked about various ways in which the study of measures in MOOC learning can inform those in face-to-face instruction learning and vice versa.

- Wilde, Zaluska & Millard (2015) [What is success anyway? Defining success in FutureLearn MOOCs](#), FutureLearn Academic Network (FLAN), 2 December, Southampton, UK.
- Wilde (2015) [What are the measurable factors for learning success that are common to face-to-face instruction and MOOCs?](#) (Lightning talk). In the *Learning Analytics LACE SoLAR Flare* networking event. 9 October, Milton Keynes, UK.
- Wilde, Zaluska & Millard (2015) [Student Success on Face-to-Face Instruction and MOOCs: What can Learning Analytics uncover?](#) In the *Web Science Education Workshop* at the ACM Web Science Conference, 28 June, Oxford, UK.

### On characterising MOOC learners

In these presentations I talked about my earlier approaches to characterising learning activity, based on unsupervised learning or statistical methods, in the Understanding Language MOOC. Here I also list a co-authored conference paper, that was an output of a Web Science Institute pump-priming project on the MOOC Observatory dashboard in which I was a co-investigator (with Su White as the principal investigator). My contribution in this paper was on the advantages and disadvantages of aggregating demographic data as well as activity data from several runs of various University of Southampton MOOCs.

- Wilde (2018) Clustering of learners' behaviour in the Understanding Language MOOC. FutureLearn Academic Network (FLAN), 7 September, Glasgow, UK.
- Wilde, León & Borthwick (2017) [Understanding Language: Understanding MOOC learners](#). In *Innovative Language Teaching and Learning at University (InnoConf17)*, 16 June, at the Centre for Research in Education and Educational Technology (CREET), the Open University, Milton Keynes, UK.
- Wilde, León & White (2016) [Tracking collective learner footprints: Aggregate analysis of MOOC learner demographics and activity](#). In the 9<sup>th</sup> Annual Int. Conf. of Education, Research and Innovation (iCERI 2016), 14-16 November, Seville, Spain.

### On using PeerWise as a peer-supported environment in HCI

In these contributions my focus was on the use of PeerWise as a peer-supported digital environment for teaching Interaction Design.

- Wilde (2019) [Rising to Challenges in Assessment and Feedback in HCI Education: A-Peer-Supported Approach](#). WAIS research group seminar, University of Southampton, 17 October.

## 7.4 Limitations

The platform-agnostic model of learner engagement within peer-supported environments considers communicative learning activities as if they would take place instantaneously, rather than over a period of time, as explained in Section 4.5. This assumption was convenient, making the model sufficiently simple, and did not cause any problems with the platforms studied, as in these datasets there were only single timestamps associated to communicative activities. However, this may not be the case for all platforms, and therefore the model would have to be extended to incorporate interval logic into communicative activities (as it is currently the case for non-communicative activities).

Another limitation of this research is that it only uses FutureLearn MOOCs for validating the model in online courses. Courses in this platform are x-MOOCs, yet follow a conversational framework, within a social-constructivist pedagogy. This may be the reason why dialogic features were especially good for characterising engagement in this courses, since these features operationalise engagement in communicative activities. Other courses for which non-communicative activities may be more dominant (such as those following a cognitive-behaviourist pedagogy) were not studied. Though one might hypothesise that the model would still be useful in characterising engagement in these contexts (and interval features be more explanatory of learner behaviour), this is something that was not explored in this research.

Similarly, another limitation of this research is that it only considers engagement in face-to-face courses that used PeerWise as a peer-supported digital environment. It is not known whether learners engaged in other platforms, such as forums in Virtual Learning Environments (VLEs), would behave in the manner here identified. In VLEs in general the focus of activity is not communicative (if seeing engagement on the platform as a whole rather than just on the forums), so it would be expected that non-communicative features would be much more explanatory of the engagement. However, this was not possible to explore in my research.

Finally, in comparing my findings against existing approaches in the literature, I focused on the heuristic-based work by [Chua et al. \(2017\)](#). I made this choice despite having identified others that have been more widely read or more influential, such as [Kizilcec et al. \(2013\)](#) and [Ferguson and Clow \(2015b\)](#). The reason I chose to do so is because in my model of learner interactions I sought to include both communicative and non-communicative activities. In particular, for the former I intended to model the turn-taking nature of communicative activities, which is well-captured in the dialogic heuristic by [Chua et al. \(2017\)](#). This approach is relatively uncharted, in comparison with attrition prediction, which had been the focus of much research around MOOCs

(including some of my own, such as in [Cobos et al. \(2017\)](#), [Ballesteros-Mesa and Wilde \(2016a\)](#), [Ballesteros-Mesa and Wilde \(2016b\)](#) and [Wilde \(2016\)](#)). Besides, attrition is not a problem of particular importance in the face-to-face instruction context. In making this choice, I was able to make meaningful comparisons of behaviours of both MOOC learners and face-to-face learners, but it is a limitation nonetheless, which needs to be taken into account when evaluating this research.

## 7.5 Future work

The following subsections describe some additional work that escapes the scope of this thesis but are directions worthwhile investigating to extend this research. I also make reference to some evidence of the viability of the studies, given preliminary experimentation in these directions (in Appendices [H](#) and [I](#)). In particular,

- Investigate further on interval features (Section [7.5.1](#));
- Perform a Principal Component Analysis (PCA) to explore the importance of all engineered features across the datasets (Section [7.5.2](#)); and,
- Investigate the adoption of PeerWise in a face-to-face learning environment when the method of assessment does not include multiple-choice questions (Section [7.5.3](#)).

### 7.5.1 Interval features

One of the feature sets that were engineered from the FutureLearn data, which correspond with an important part of the platform-agnostic model described in Chapter [4](#) were those related to algebra of intervals. Given that features of these kind are likely to be found in many peer-supported digital environments, it is of interest to explore how informative features from this class are in categorising learner engagement. Indeed, as part of my doctoral research, I conducted some exploratory experiments and conducted some preliminary analysis on MOOC data using these features, as shown in Appendix [H](#). However, I chose to exclude these findings from the main analysis on MOOC data in Chapter [5](#) because these features were not extractable from the PeerWise dataset that was made available to me at the time of writing.

Having excluded it, however, interval information is captured in the system, namely defined by the timestamps when students starts contributing to a question and when

they submit them (as per a recent communication with Paul Denny<sup>2</sup>). Therefore it is feasible to extend this research in the near future by using those features in the analysis, once the extended dataset about the PeerWise data become available.

## 7.5.2 Principal Component Analyses

An important part of the research presented in this thesis involved feature engineering and through this process, I defined 78 features for MOOC data and 72 features for PeerWise data. Several design decisions subsequently drove me to select several (small) subsets of these features for the analyses presented, which were informed by domain knowledge and the focus on achieving interpretable results for practice. Another important factor in selecting the reduced feature sets was need to look at a common operationalisation of learner engagement across peer-supported digital environments (as per the aim of this research) rather than a data-driven, systematic study of the fitness of all the features available about learners in each of these environments.

A Principal Component Analysis (PCA) involving a much larger set from the engineered and extracted features per course could allow for the identification of critical discriminators not in the categorisations presented in this thesis. A PCA involving all of the features was proven to not account for a reasonable proportion of variability in the data, and therefore not considered it would provide a robust clustering. However, perhaps a systematic inclusion of certain features at a time would provide factors that include information about learner engagement in non-communicative activity in a reduced feature space that keeps a reasonable variability. This is something that could be explored.

An interesting direction of work would begin with performing a series of comparative PCAs systematically over all the courses in consideration allowing us to assert what features appear consistently high in importance across all courses (both in MOOCs and PeerWise). Subsequently, I would investigate whether the inclusion of additional features (present in at least one dataset but not all) affect the clustering performance as presented in this thesis.

---

<sup>2</sup>In subsequent email exchanges with Paul Denny, dated 14 December 2020, I learned that “PeerWise maintains detailed log files which record all kinds of interactions. For example, it would be possible to measure when a student begins creating a question and when that question is actually published. Or, when a student looks at a question and then when they submit an answer to that question.” Studying some of that data, by engineering features using the model would be interesting future work.

### 7.5.3 PeerWise adoption when not aligned with assessment

Despite participation not being rewarded with marks, the second cohort using PeerWise for COMP2213 did engage. Albeit in smaller numbers, it was still a significant number of students (62%). Many fewer questions were created but a large number of answers were provided, evidence that students used the tool as a revision aid. One of the reasons to justify that level of voluntary engagement is that the final exam (worth 50% of the module) included a large component of multiple-choice questions.

However, it would be interesting to study the participation should that component be removed, and I have ethical approval to do so with two of my modules in 2020/21 at the University of Winchester. These two modules are: BS1912 Information Systems and Organisation (for first year undergraduate students) and BS2203 Secure Systems Architecture (for third years). In both modules the method of assessment was an end-of-semester report worth 100% of the marks. The method of delivery was hybrid learning (a combination of face-to-face and synchronised online learning, according to a rota). The rota was designed such that class sizes were no more than ten students at a time in face-to-face lectures.

I hypothesise that this misalignment with the assessment method has an effect on strategic learners, who then would be less inclined to engage in PeerWise. However, it might provide a much-needed additional way of receiving peer-support and engaging with each other, particularly for the first-year cohort, given that they had not yet had the opportunity to create bonds in face-to-face interactions as they started their studies with strict social distancing measures in place at the university.

## 7.6 Concluding remarks

At the start of my PhD journey, I set out to improve an understanding of how the underlying learning phenomena are manifested, both in face-to-face instruction and in MOOCs. The work I presented in this thesis goes a significant distance to achieve this. My model of learner engagement within peer-supported digital environments provides a solid theoretical framework to analyse digital traces of engagement in both fully online courses and face-to-face courses complemented with digital environments. Further, I built on state-of-the-art knowledge of classification of MOOC learners using unsupervised learning models, clustering in particular, with a feature engineering process that was guided by the model. I subsequently was able to apply the same process to data from PeerWise, obtaining interpretable clusters of learning engagement for each platform. This enabled



me to do a meaningful comparison of this phenomenon as observed in both platforms, overcoming the challenges of heterogeneity in the data representation of their respective captured traces of engagement, and finding that once behavioural constraints in relation to assessment incentives were removed from my face-to-face instruction course, learners behaved in a very similar way to those in MOOCs.

My hope is that this work contributes to a greater understanding of how learners engage in these platforms, an understanding that would better inform learning design and interventions that ultimately support learners in a significantly positive way: fashioning stepping stones out of stumbling blocks.



# Bibliography

- Alessandro Acquisti. **Nudging privacy: The behavioral economics of personal information.** *IEEE Security & Privacy*, 7(6):82–85, 2009.
- Carlos Alario-Hoyos, Mar Pérez-Sanagustín, Carlos Delgado-Kloos, Hugo A. Parada G., and Mario Muñoz-Organero. **Delving into participants' profiles and use of social tools in MOOCs.** *IEEE Transactions on Learning Technologies*, 7(3):260–266, 2014.
- James F Allen. **Maintaining knowledge about temporal intervals.** *Communications of the ACM*, 26(11):832–843, 1983.
- Abram Anders. **Theories and applications of massive online open courses (MOOCs): The case for hybrid design.** *The International Review of Research in Open and Distributed Learning*, 16(6), 2015.
- Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. **Engaging with massive online courses.** In *Proceedings of the 23<sup>rd</sup> International Conference on World Wide Web, WWW '14*, page 687–698, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327442.
- Paul Baepler and Cynthia James Murdoch. **Academic Analytics and Data Mining in Higher Education.** *International Journal for the Scholarship of Teaching and Learning*, 4(2), July 2010.
- Ryan Baker. **Feature engineering: Better, more interpretable models.** In *Learning Analytics Learning Network (LALN)*, August 2020.
- Rebecca Balebako, Pedro G. León, Hazim Almuhammedi, Patrick Gage Kelley, Jonathan Mugan, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. **Nudging users towards privacy on mobile devices.** In *ACM Workshop of the Conference on Human Factors in Computing Systems (CHI 2011)*, 2011.

- Miguel Ballesteros-Mesa and Adriana Wilde. **Data analytics for MOOC providers**. In Rosabel Roig-Vila, editor, *XIX Congreso Internacional EDUTEC 2016*, Alicante, Spain, November 2016a. University of Alicante, Octaedro.
- Miguel Ballesteros-Mesa and Adriana Wilde. **Recommendations arising from performing data analytics on FutureLearn courses**. In *FutureLearn Academic Network (FLAN) Meeting*, Leicester, United Kingdom, November 2016b. University of Leicester.
- Rebecca Barber and Mike Sharkey. **Course correction: using analytics to predict course success**. In *Proceedings of the 2<sup>nd</sup> International Conference on Learning Analytics and Knowledge*, pages 259–262. ACM, 2012.
- Simon P Bates, Ross K Galloway, and Karon L McBride. **Student-generated content: Using PeerWise to enhance engagement and outcomes in introductory physics courses**. In *AIP Conference Proceedings*, volume 1413, pages 123–126. AIP, 2012.
- David Biggins, Emma J Crowley, Elvira Bolat, Mihai Dupac, and H Dogan. **Using PeerWise to improve engagement and learning**. In *The European Conference on Education*. The International Academic Forum (IAFOR), July 2015.
- John Biggs and Catherine Tang. *Teaching for Quality Learning at University*. Open University Press, third edition, 2007.
- Alejandro Bogarín, Cristóbal Romero, Rebeca Cerezo, and Miguel Sánchez-Santillán. **Clustering for improving educational process mining**. In *Proceedings of the fourth international conference on learning analytics and knowledge*, pages 11–15, 2014.
- Pascal Bruegger, Adriana Wilde, and Loic Guibert. **On the development of a resident monitoring system: Usability, privacy and security aspects**. In *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, pages 288–293, 2020.
- Bradley Carron-Arthur, John A Cunningham, and Kathleen M Griffiths. **Describing the distribution of engagement in an internet support group by post frequency: A comparison of the 90-9-1 principle and Zipf's law**. *Internet Interventions*, 1(4):165–168, 2014.
- Shi Min Chua, Caroline Tagg, Mike Sharples, and Bart Rienties. **Discussion analytics: Identifying conversations and social learners in futurelearn MOOCs**. In Lorenzo Vigenini, Yuan (Elle) Wang, Luc Paquette, and Manuel León Urrutia, editors, *FutureLearn*

- data: what we currently have, what we are learning and how it is demonstrating learning in MOOCs*, volume 1967 of *CEUR-WS.org*, pages 74–93. Simon Fraser University, March 2017.
- Doug Clow. **MOOCs and the funnel of participation**. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 185–189, 2013.
- Ruth Cobos, Adriana Wilde, and Ed Zaluska. **Predicting attrition from Massive Open Online Courses in FutureLearn and edX**. In Lorenzo Vigentini, Yuan (Elle) Wang, Luc Paquette, and Manuel León Urrutia, editors, *FutureLearn data: what we currently have, what we are learning and how it is demonstrating learning in MOOCs*, volume 1967 of *CEUR-WS*, pages 74–93, Vancouver, Canada, March 2017. Simon Fraser University, Learning Analytics and Knowledge (LAK'17).
- Louis Cohen, Lawrence Manion, and Keith Morrison. *Research methods in education*. Routledge, sixth edition, 2007. ISBN 13:978-0-415-36878.
- Linda Darling-Hammond, Lisa Flook, Channa Cook-Harvey, Brigid Barron, and David Osher. **Implications for educational practice of the science of learning and development**. *Applied Developmental Science*, 24(2):97–140, 2020.
- Gareth R. Davies, Hereward Proops, and Clare M. Carolan. **The development and use of a multiple-choice question (MCQ) assessment to foster deeper learning: An exploratory web-based qualitative investigation**. *Journal of Teaching and Learning Special Issue: Digital Learning in Higher Education*, 14(1):1–12, 2020.
- Paul Denny. **The effect of virtual achievements on student engagement**. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 763–772, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990.
- Paul Denny, John Hamer, Andrew Luxton-Reilly, and Helen Purchase. **PeerWise: students sharing their multiple choice questions**. In *Proceedings of the fourth international workshop on Computing Education Research*, pages 51–58. ACM, 2008a.
- Paul Denny, Andrew Luxton-Reilly, and John Hamer. **The PeerWise system of student contributed assessment questions**. In *Proceedings of the tenth conference on Australasian computing education-Volume 78*, pages 69–74. Australian Computer Society, Inc., 2008b.
- Paul Denny, Ewan Tempero, Dawn Garbett, and Andrew Petersen. **Examining a student-generated question activity using random topic assignment**. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE

- '17, pages 146–151, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450347044.
- Michael Derntl. *Patterns for Person-Centered E-learning*. PhD thesis, University of Vienna, 2005. <http://elearn.pri.univie.ac.at/derntl/diss/diss-derntl.pdf>.
- Jim Devon, James H Paterson, David C Moffat, and June McCrae. **Evaluation of student engagement with peer feedback based on student-generated MCQs**. *Innovation in Teaching and Learning in Information and Computer Sciences*, 11(1):27–37, 2012.
- Alan Dix. **Challenge and potential of fine grain, cross-institutional learning data**. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale, L@S '16*, pages 261–264, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450337267.
- Nia Dowell and Oleksandra Poquet. **SCIP: Combining group communication and interpersonal positioning to identify emergent roles in scaled digital environments**. *Computers in Human Behavior*, 2021.
- Nia Dowell, Oleksandra Poquet, and Christopher Brooks. **Applying group communication analysis to educational discourse interactions at scale**. In *ICLS 2018 Proceedings*, pages 1815–1822. International Society of the Learning Sciences, Inc.[ISLS], 2018.
- Elaine Doyle and Patrick Buckley. **The impact of co-creation: an analysis of the effectiveness of student authored multiple choice questions on achievement of learning outcomes**. *Interactive Learning Environments*, 0(0):1–10, 2020.
- Stephen W Draper. **Catalytic assessment: understanding how MCQs and EVS can foster deep learning**. *British Journal of Educational Technology*, 40(2):285–293, 2009.
- Ismail Duru, Ayse Saliha Sunar, Su White, and Banu Diri. **Deep learning for discussion-based cross-domain performance prediction of MOOC learners grouped by language on FutureLearn**. *Arabian Journal for Science and Engineering*, 2021.
- Ashish Dutt, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, and Hamidreza Mahroeian. **Clustering algorithms applied in educational data mining**. *International Journal of Information and Electronics Engineering*, 5(2):112–116, March 2015.
- Tom Fawcett. **An introduction to ROC analysis**. *Pattern Recognition Letters*, 27(8):861–874, June 2006. ISSN 0167-8655.
- Rebecca Ferguson and Doug Clow. **Consistent commitment: Patterns of engagement across time in massive open online courses (MOOCs)**. *Journal of Learning Analytics*, 2(3):55–80, 2015a.

- Rebecca Ferguson and Doug Clow. **Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs)**. In *Proceedings of the fifth international conference on learning analytics and knowledge (LAK'15)*, pages 51–58, 2015b.
- Antonio Fini, Andreas Formiconi, Alessandro Giorni, Nuccia Pirruccello, Elisa Spadavecchia, and Emanuela Zibordi. **IntroOpenEd 2007: An experience on open education by a virtual community of teachers**. *Journal of e-Learning and Knowledge Society*, 4(1):231–239, 2008.
- Richard Fox. **Constructivism examined**. *Oxford Review of Education*, 27(1):23–35, 2001.
- William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. **Knowledge discovery in databases: An overview**. *AI Magazine*, 13(3):57, 1992.
- Paulo Freire. *Pedagogía del oprimido*. Tierra Nueva, Montevideo, 1970.
- Nabeel Gillani and Rebecca Eynon. **Communication patterns in massively open online courses**. *The Internet and Higher Education*, 23:18–26, 2014.
- Joseph G Glynn, Paul L Sauer, and Thomas E Miller. **Signaling student retention with prematriculation data**. *Journal of Student Affairs Research and Practice (NASPA, National Association of Student Personnel Administrators)*, 41(1), December 2003.
- Robert Godwin-Jones. **Challenging hegemonies in online learning**. *Language Learning & Technology*, 16(2):4–13, 2012.
- Hannah Gore. *Engagement of Learners Undertaking Massive Open Online Courses and the Impact of Design*. PhD thesis, The Open University., 2018.
- Great Britain. Parliament. House of Lords. **Science and Technology Committee - Second Report Behaviour Change**. Technical Report (HL 2010-12 (179), Great Britain and Parliament and “House of Commons”, July 2011. Chapter 2: Definitions, Categorisation and the Ethics of Behaviour Change Interventions.
- Nicky Hockly. **The digital generation**. *ELT Journal*, 65(3):322–325, 2011.
- Gwyneth Hughes. **Diversity, Identity and Belonging in e-Learning Communities: Some Theories and Paradoxes**. *Teaching in Higher Education*, 12(5-6):709–720, 2007.
- Louise Humpage et al. **PeerWise: A useful learning tool for sociology?** *New Zealand Sociology*, 29(1):135, 2014.
- Karen Hunsdale, Dineshen Chuckravanen, Jacqueline Daykin, and Amar Seeam. **Allen’s interval algebra and smart-type environments**. *International Journal on Advances in Software*, 2017. ISSN 19422628.

- Tomi Janhunnen and Michael Sioutis. **Allen's interval algebra makes the difference**. In Petra Hofstedt, Salvador Abreu, Ulrich John, Herbert Kuchen, and Dietmar Seipel, editors, *Declarative Programming and Knowledge Management*, pages 89–98, Cham, 2020. Springer International Publishing. ISBN 978-3-030-46714-2.
- Christopher Jones. Students, the net generation, and digital natives. In Michael Thomas, editor, *Deconstructing Digital Natives: Young People, Technology, and the New Literacies*, pages 30–45. Taylor & Francis, 2011a.
- David Thomas Jones. *An Information Systems Design Theory for e-Learning*. PhD thesis, Australian National University, February 2011b. <http://hdl.handle.net/1885/8370>.
- David Kember. **A Reconceptualisation of the Research into University Academics' Conceptions of Teaching**. *Learning and instruction*, 7(3):255–275, 1997.
- Gregor Kennedy, Terry Judd, Barney Dalgarno, and Jenny Waycott. **Beyond natives and immigrants: exploring types of net generation students**. *Journal of Computer Assisted Learning*, 26(5):332–343, 2010.
- René F Kizilcec and Maximillian Chen. **Student engagement in mobile learning via text message**. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, pages 157–166, 2020.
- René F Kizilcec, Chris Piech, and Emily Schneider. **Deconstructing disengagement: analyzing learner subpopulations in massive open online courses**. In *Proceedings of the third international conference on Learning Analytics and Knowledge, LAK'13*, pages 170–179. ACM, 2013.
- David Kolb. Experiential learning: From discourse model to conversation. *Lifelong Learning in Europe*, 3:148–153, 01 1998.
- Agnes Kukulska-Hulme, Carina Bossu, Tim Coughlan, Rebecca Ferguson, Elizabeth FitzGerald, Mark Gaved, Cristothea Herodotou, Bart Rienties, Julia Sargent, Eileen Scanlon, Jinlan Tang, Qi Wang, Denise Whitelock, and Shuai Zhang. **Innovating pedagogy 2021: Open university innovation report 9**, 2021.
- Diana Laurillard. *Rethinking university teaching: A conversational framework for the effective use of learning technologies*. Routledge, 2013.
- Allison Littlejohn, Colin Milligan, and Anoush Margarayn. **Collective learning in the workplace: Important knowledge sharing behaviours**. *International Journal of Advanced Corporate Learning (iJAC)*, 4(4):26–31, November 2011. ISSN 1867-5565.



- Ren Liu and Kenneth Koedinger. **Going beyond better data prediction to create explanatory models of educational data.** In Charles Lang, George Siemens, Alyssa Friend Wise, and Dragan Gašević, editors, *The Handbook of Learning Analytics*, pages 69–76. Society for Learning Analytics Research (SoLAR), Alberta, Canada, 1 edition, 2017. ISBN 978-0-9952408-0-3.
- Andrew Luxton-Reilly, Paul Denny, Beryl Plimmer, and Robert Sheehan. **Activities, affordances and attitude: How student-generated questions assist learning.** In *Proceedings of the 17th ACM Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE '12*, pages 4–9, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312462.
- Aaron Mac Raighne, Morag Casey, Robert Howard, and Barry Ryan. **Student attitudes to an online, peer-instruction, revision aid in science education.** *Journal of Perspectives in Applied Academic Practice*, 3:49–60, 2015.
- Mirna Carelli Oliveira Maia, Eliane Cristina Araújo, Jorge Figueiredo, and Dalton Serey. **Student engagement through creation of new activities: An empirical study on contributing student pedagogy.** In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1693–1702. SBC, 2020.
- Alexander McAuley, Bonnie Stewart, George Siemens, Dave Cormier, and Creative Commons. **The MOOC model for digital practice.** Technical report, University of Prince Edward Island, Charlottetown, Canada, 2010.
- Stephen McClean. **Implementing PeerWise to engage students in collaborative learning.** *Perspectives on Pedagogy and Practice*, 6:89–96, 2015.
- Jeff Mehring. **Present research on the flipped classroom and potential tools for the efl classroom.** *Computers in the Schools*, 33(1):1–10, 2016.
- Colin Milligan, Allison Littlejohn, and Anoush Margarayn. **Patterns of engagement in connectivist MOOCs.** *Journal of Online Learning and Teaching*, 9(2):149–159, 2013.
- David Nicol. **E-assessment by design: using multiple-choice tests to good effect.** *Journal of Further and Higher Education*, 31(1):53–64, 2007.
- Olivia Ojuroye, Russel Torah, Steve Beeby, and Adriana Wilde. **Smart textiles for smart home control and enriching future wireless sensor network data.** In *Sensors for Everyday Life*, pages 159–183. Springer, 2017.
- John Palfrey and Urs Gasser. *Born Digital: Understanding the First Generation of Digital Natives.* Basic Books (AZ), 2010.

- Laura Pappano. **The Year of the MOOC**. *The New York Times*, pages 1–7, 2012. ISSN 0362-4331.
- Luc Paquette and Ryan S Baker. **Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system**. *Interactive Learning Environments*, 27(5-6):585–597, 2019.
- Gordon Pask. *The cybernetics of human learning and performance: A guide to theory and research*. Hutchinson & Co, 1975. ISBN 0-09-119490-3.
- Dan Pelleg and Andrew Moore. **X-means: Extending K-means with efficient estimation of the number of clusters**. In *Proceedings of the 17<sup>th</sup> international conference on machine learning (ICML)*, pages 277–281, 2000.
- Vladimir Pestov. **Is the k-NN classifier in high dimensions affected by the curse of dimensionality?** *Computers & Mathematics with Applications*, 65(10):1427–1437, 2013. ISSN 0898-1221. Grasping Complexity.
- Pireh Pirzada, Neil White, and Adriana Wilde. **Sensors in smart homes for independent living of the elderly**. In *5<sup>th</sup> International Multi-Topic ICT Conference (IMTIC)*, pages 1–8, 2018.
- Pireh Pirzada, Adriana Gabriela Wilde, Gayle Helane Doherty, and David Harris-Birtill. **Ethics and acceptance of smart homes for older adults**. *Informatics for Health and Social Care*, April 2021.
- Oleksandra Poquet, Jelena Jovanovic, and Shane Dawson. **Differences in forum communication of residents and visitors in MOOCs**. *Computers & Education*, 156:103937, 2020. ISSN 0360-1315.
- Marc Prensky. **Digital Natives, Digital Immigrants Part 2: Do They Really Think Differently?** *On the Horizon*, 9(6):1–9, 2001.
- Justin Reich and José A. Ruipérez-Valiente. **The MOOC pivot**. *Science*, 363(6423):130–131, 2019. ISSN 0036-8075.
- Adrian Renzo et al. **Multiple-choice questions in the humanities: a case study of PeerWise in a first-year popular music course**. In *Rhetoric and Reality: Critical perspectives on educational technology*, pages 506–564. Dunedin, New Zealand: Proceedings of ASCILITE 2014 - 31<sup>st</sup> Annual Conference of the Australian Society for Computers in Tertiary Education 2014, November 2014.
- Amy L. Reschly and Sandra L. Christenson. **Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct**. In *Handbook of research on student engagement*, chapter 1, pages 3–19. Springer, 2012.

- Johanna Rhodes. **Using PeerWise to knowledge build and consolidate knowledge in nursing education.** *Southern Institute of Technology Journal of Applied Research (SIT-JAR)*, 2013.
- Johanna Rhodes et al. **Using PeerWise in nursing education-a replicated quantitative descriptive research study.** *Kai Tiaki Nursing Research*, 6(1):10, 2015.
- C Osvaldo Rodriguez. **MOOCs and the AI-Stanford like courses: Two successful and distinct course formats for massive open online courses.** *European Journal of Open, Distance and E-Learning*, page 13, 2012. ISSN ISSN-1027-5207.
- Cristóbal Romero, Rebeca Cerezo, Alejandro Bogarín, and Miguel Sánchez-Santillán. **Educational process mining: a tutorial and case study using Moodle data sets.** *Data mining and learning analytics: Applications in educational research*, pages 1–28, 2016.
- Cristóbal Romero and Sebastián Ventura. **Educational data mining: a review of the state of the art.** *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618, 2010.
- Neil Rubens, Dain Kaplan, and Toshio Okamoto. **E-Learning 3.0: anyone, anywhere, anytime, and AI.** In *New Horizons in Web Based Learning*, pages 171–180. Springer, 2014.
- Gilly Salmon. *E-tivities: The key to active online learning*. Taylor and Francis, 2002.
- Farhana Sarker. *Linked Data Technologies to Support Higher Education Challenges: Student Retention, Progression and Completion*. PhD thesis, Electronics and Computer Science, University of Southampton, 2014.
- Karen Scouller. **The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay.** *Higher Education*, 35(4):453–472, 1998.
- Dhawal Shah. **By the numbers: MOOCs in 2020 (analysis).** *The Report by Class Central*, November 2020a.
- Dhawal Shah. **The second year of the MOOC: A review of MOOC stats and trends in 2020 (analysis).** *The Report by Class Central*, December 2020b.
- Mike Sharkey. **Academic analytics landscape at the University of Phoenix.** In *Proceedings of the 1<sup>st</sup> International Conference on Learning Analytics and Knowledge (LAK)*, pages 122–126. ACM, 2011.

- Mike Sharples and Rebecca Ferguson. **Pedagogy-informed design of conversational learning at scale**. In *CEUR Workshop Proceedings*, volume 2437, page 14, 2019. EC-TEL Practitioner Proceedings: 14<sup>th</sup> European Conference on Technology-Enhanced Learning.
- George Siemens. **Connectivism: a learning theory for the digital age**. *International Journal of Instructional Technology and Distance Learning*, 2(1), 2005.
- George Siemens and Ryan SJD Baker. **Learning analytics and educational data mining: towards communication and collaboration**. In *Proceedings of the 2<sup>nd</sup> international conference on Learning Analytics and Knowledge (LAK)*, pages 252–254. ACM, 2012.
- George Siemens and Phil Long. **Penetrating the fog: Analytics in learning and education**. *Educause Review*, 46(5):30–32, 2011.
- Adalberto L Simeone, Marco Speicher, Andreea Molnar, Adriana Wilde, and Florian Daiber. **LIVE: The human role in Learning in Immersive Virtual Environments**. In *Symposium on Spatial User Interaction, SUI '19*, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369756.
- Sharon Slade and Paul Prinsloo. **Learning analytics: Ethical issues and dilemmas**. *American Behavioral Scientist*, 57(10):1510–1529, 2013.
- Stephen Snow and Adriana Wilde. **Supporting authoring of multiple-choice questions in human-computer interaction using PeerWise**. In Carol Evans, editor, *What Works in Assessment and Feedback: Simply Better conference*, Southampton, UK, September 2017. University of Southampton.
- Stephen Snow, Adriana Wilde, Paul Denny, and m.c. schraefel. **A discursive question: Supporting student-authored multiple choice questions through peer-learning software in non-STEMM disciplines**. *British Journal of Educational Technology, BJET*, pages 1–16, 2018.
- Pei-Chen Sun, Ray J Tsai, Glenn Finger, Yueh-Yang Chen, and Dowming Yeh. **What drives a successful e-learning? an empirical investigation of the critical factors influencing learner satisfaction**. *Computers & Education*, 50(4):1183–1202, 2008.
- Ayse Saliha Sunar, Rabeeh Ayaz Abbasi, Hugh C. Davis, Su White, and Naif R. Aljohani. **Modelling MOOC learners' social behaviours**. *Computers in Human Behaviour*, 107 (105835):1–12, 2020.
- Ayse Saliha Sunar, Su White, Nor Aniza Abdullah, and Hugh C. Davis. **How learners' interactions sustain engagement: A MOOC case study**. *IEEE Transactions on Learning Technologies*, 10(4):475–487, 2017.

- Richard Thaler and Cass Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008.
- Jo Tondeur, Johan van Braak, Fazilat Siddiq, and Ronny Scherer. **Time for a new approach to prepare future teachers for educational technology use: Its meaning and measurement**. *Computers & Education*, 94:134–150, 2016.
- Shu-Fen Tseng, Yen-Wei Tsao, Liang-Chih Yu, Chien-Lung Chan, and K. Robert Lai. **Who will pass? analyzing learner behaviors inMOOCs**. *Research and Practice in Technology Enhanced Learning*, 11(1):8, 2016.
- Philip Tubman, Murat Oztok, and Phil Benachour. **New platform affordances for encouraging social interaction in MOOCs: The “comment discovery tool” interactive visualisation plugin**. In *2019 IEEE 19<sup>th</sup> International Conference on Advanced Learning Technologies (ICALT)*, volume 2161-377X, pages 34–36, July 2019.
- Sofia Loredana Tudor. **Formal – non-formal – informal in education**. *Procedia - Social and Behavioral Sciences*, 76:821–826, 2013. ISSN 1877-0428. 5<sup>th</sup> International Conference EDU-WORLD 2012 - Education Facing Contemporary World Issues.
- Tiffany Unruh, Michelle L Peters, and Jana Willis. **Flip this classroom: A comparative study**. *Computers in the Schools*, 33(1):38–58, 2016.
- Trevor van Mierlo, Sabrina Voci, Sharon Lee, Rachel Fournier, and Peter Selby. **Super-users in social networks for smoking cessation: analysis of demographic characteristics and posting behavior from the canadian cancer society’s smokers’ helpline online and stopsmokingcenter.net**. *Journal of medical Internet research*, 14(3):e66, 2012.
- Anna Vasilchenko, Adriana Wilde, Stephen Snow, Madeline Balaam, and Marie Devlin. **Video coursework: Opportunity and challenge for HCI education**. In Tiziana Catarci, Kent Norman, and Massimo Mecella, editors, *Proceedings of the 2018 ACM International Conference on Advanced Visual Interfaces*, Castiglione della Pescaia, Grosseto, Italy, May 2018. ACM.
- Tracey Walker. **A research-based insight into FutureLearners. part 1: What we did and why**. *FutureLearn Newsletter*, January 2018a. As reported by Niamh O’Grady.
- Tracey Walker. **A research-based insight into FutureLearners. part 2: The ‘Work and Study’ archetypes**. *FutureLearn Newsletter*, February 2018b. As reported by Niamh O’Grady.
- Tracey Walker. **A research-based insight into FutureLearners. part 3: The ‘Personal Life’ archetypes**. *FutureLearn Newsletter*, February 2018c. As reported by Niamh O’Grady.

- Tracey Walker. **A research-based insight into FutureLearners. part 4: The ‘Leisure’ archetypes.** *FutureLearn Newsletter*, February 2018d. As reported by Niamh O’Grady.
- Jing Wang. Sentiment analysis of comments in Massive Open Online Courses. A Dissertation for the degree of MSc Data Science, University of Southampton, Southampton, United Kingdom, September 2017. Supervised by Adriana Wilde.
- David S. White, Lynn Silipigni Connaway, Donna Lanclos, Alison Le Cornu, and Erin Hood. **Digital visitors and residents: Progress report.** *Joint Information Systems Committee (JISC), University of Oxford, OCLC, University of North Carolina*, June 2012.
- Adriana Wilde. **Adapting to class sizes: what feedback fits best?** In *Improving student satisfaction with assessment and feedback – one day conference*, University of Southampton, February 2014.
- Adriana Wilde. **Student smartphones: tools or barriers? attitudes amongst students in higher education in Chile and the UK.** In *womENCourage*, Uppsala, Sweden, September 2015a. ACM.
- Adriana Wilde. **What are the measurable factors for learning success that are common to face-to-face instruction and MOOCs.** In *Learning Analytics LACE SoLAR Flare event*, Milton Keynes, United Kingdom, October 2015b. The Open University.
- Adriana Wilde. **Understanding persuasive technologies to improve completion rates in MOOCs.** In *Proceedings of the 2016 International Conference on Advanced Visual Interfaces (AVI’16)*, Bari , Italy, June 2016. ACM.
- Adriana Wilde. **Rising to challenges in assessment and feedback in HCI education: A-peer-supported approach.** Seminar for the Web and Internet Science research group, University of Southampton, October 2019.
- Adriana Wilde. **Choosing between wider participation and quality of interactions: a study of learner engagement within PeerWise.** In *The United Kingdom and Ireland Computing Education Research conference (UKICER)*, Glasgow, United Kingdom, September 2020. online.
- Adriana Wilde. **Feature engineering for clustering analysis of large and heterogeneous educational datasets.** In *the 5<sup>th</sup> Women in Data Science (WiDS) Cambridge conference*, Virtual event, Cambridge, Massachussetts, March 2021.
- Adriana Wilde, Miguel Ballesteros-Mesa, and Manuel León Urrutia. **Hacia un marco de análisis del aprendizaje en cursos en línea masivos y abiertos: informando al proveedor.** In Rosabel Roig-Vila, editor, *EDUcación y TECnología. Propuestas desde la investigación y la innovación educativa*, pages 276–277. Octaedro, 2016a.



- Adriana Wilde and Alan Dix. **Navigating challenges on wide-scale adoption of video for HCI education: the HCIvideoW experience.** In *2020 ACM Learning at Scale Conference, COVID-19 case study track*, Virtual event, August 2020a.
- Adriana Wilde and Alan Dix. **Second workshop on using video in computer science education.** In *ACM UK and Ireland Computer Science Education Conference (UKICSE)*, Virtual event, September 2020b.
- Adriana Wilde, Alan Dix, Chris Evans, Anna Vasilchenko, Joseph Maguire, and Stephen Snow. **Towards a taxonomy of video for HCI education.** In *Trends and good practices in research and teaching: a Spanish-English collaboration*, pages 31–46. Octaedro, 2019.
- Adriana Wilde, Manuel León Urrutia, and Su White. **Tracking collective learner footprints: Aggregate analysis of mooc learner demographics and activity.** In *ICERI2016 Proceedings*, 9<sup>th</sup> annual International Conference of Education, Research and Innovation, pages 1404–1413. IATED, 14-16 November, 2016 2016b. ISBN 978-84-617-5895-1.
- Adriana Wilde, Olivia Ojuroye, and Russel Torah. **Prototyping a voice-controlled smart home hub wirelessly integrated with a wearable device.** In *2015 9<sup>th</sup> International Conference on Sensing Technology (ICST)*, pages 71–75, 2015.
- Adriana Wilde and Olja Rastić-Dulborough. **Encouraging gender diversity in computing by supporting women’s participation in conferences.** In *WomENCourage*, Barcelona, Spain, September 2017. ACM.
- Adriana Wilde and Stephen Snow. **Addressing challenges in assessing Human-Computer Interaction at scale.** In *Proceedings of the Computing Education Practice conference*, Durham, United Kingdom, January 2018a. University of Durham.
- Adriana Wilde and Stephen Snow. **Passive consumers no more: Experiences with students producing video materials for assessment in HCI.** In Adriana Wilde, Anna Vasilchenko, and Alan Dix, editors, *HCI and the educational technology revolution #HCIed2018: a workshop on video-making for teaching and learning human-computer interaction*, Castiglione della Pescaia, Grosseto, Italy, May 2018b.
- Adriana Wilde and Kasim Terzic. **Workshop on using video in computer science education.** In *University of St Andrews*, St Andrews, UK, August 2018.
- Adriana Wilde, Manuel León Urrutia, and Kate Borthwick. **Understanding language: understanding MOOC learners.** In Fernando Rosell-Aguilar, Tita Beaven, and Mara Fuertes-Gutiérrez, editors, *Proceedings of the 7<sup>th</sup> Annual Conference in the Innovative Language Teaching and Learning at University (InnoConf17)*, Milton Keynes, United Kingdom, June 2017.

- Adriana Wilde, Anna Vasilchenko, and Alan Dix. **HCI and the educational technology revolution #HCIED2018: A workshop on video-making for teaching and learning human-computer interaction**. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces, AVI '18*, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356169.
- Adriana Wilde and Jing Wang. **Sentiment analysis in a FutureLearn MOOC**. In *Future-Learn Academic Network (FLAN) Meeting*, London, United Kingdom, November 2017. British Council.
- Adriana Wilde and Ed Zaluska. *Tecnología Innovación e Investigación en los Procesos de Enseñanza-Aprendizaje.*, chapter Held-by-hand learners: a survey of technologies to support positive behaviours of Higher Education students today, pages 3122–3132. 10045/61787. Octaedro, Barcelona, 2016. ISBN 978-84-9921-848-9.
- Adriana Wilde, Ed Zaluska, and Dave Millard. **What is success anyway? – Defining success in FutureLearn MOOCs**. In *FutureLearn Academic Network (FLAN) Meeting*, Southampton, United Kingdom, December 2015. University of Southampton.
- Alyssa Friend Wise and Yi Cui. **Learning communities in the crowd: Characteristics of content related interactions and social relationships in MOOC discussion forums**. *Computers & Education*, 122:221 – 242, 2018. ISSN 0360-1315.
- Ian H. Witten and David Bainbridge. **A retrospective look at Greenstone: Lessons from the first decade**. In *Proceedings of the 7<sup>th</sup> ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07*, page 147–156, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936448.
- Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining. Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Elsevier, Cambridge, USA, 2017. ISBN: 978-0-12-804291-5.
- Nicolas Zurbuchen, Pascal Bruegger, and Adriana Wilde. **A comparison of machine learning algorithms for fall detection using wearable sensors**. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pages 427–431, 2020.
- Nicolas Zurbuchen, Adriana Wilde, and Pascal Bruegger. **A machine learning multi-class approach for fall detection systems based on wearable sensors with a study on sampling rates selection**. *Sensors*, 21(3), 2021. ISSN 1424-8220.



## Ethics and Research Governance Online 2 at Southampton

The following are the documents related to the ethical approval ERGO/FEPS/55694 for secondary data analysis as described in Chapters 5 and 6 on MOOCs and PeerWise respectively.

- Screenshot of the ERGO2 system showing the approval status of the final amendment of the submission (in page A-2).
- Original Ethics application for secondary data analysis, dated 2nd March 2020 (in pages A-3 to A-9).
- First amendment to the submission, dated 7th June 2020, with tracked changes visible (in pages A-10 to A-16). This amendment was prompted by the Data Protection Impact Assessment (DPIA) exercise detailed in Appendix B.
- Second amendment to the submission, dated 11th May 2021 (in pages A-17 to A-27). This amendment was prompted upon examination of the inconsistency between the DPIA and the previously approved ERGO application with regards to the declared intention to release the anonymised feature datasets. This was correctly addressed in the answer to question 12 ("How will you store and manage the data *before* and *during* the analysis? What will happen *at the end* of the project?")

You are logged in as Adriana Wilde (Log out) Accessibility Tools

# ERGO II

Ethics and Research Governance Online

UNIVERSITY OF  
**Southampton**

Home Submissions ▾

## 55694.A2 - Learner Engagement within Peer-Supported Environments mediated by Digital Technologies: PeerWise and MOOC Platforms (Amendment 2)

Submission Overview
Submission Questionnaire
Attachments
History

Details

**Status** Approved

**Category** Category B

**Submitter's Faculty** Faculty of Engineering and Physical Sciences (FEPS)

The end date for this study is currently 25 August 2021

📅 Request extension

*If you are making any other changes to your study please create an amendment using the button below.*

Latest Review Comments

14/05/2021 19:48:59 - Committee: Approved

No comments

14/06/2021 11:59:35 - Committee: Approved

Comments:

The clarifications about the way data is processed to remove potentially identifying characteristics and made suitable for submission as an open dataset is noted, and found to be sensible.

Amendment History

- 📄 **Latest Version 55694.A2** (Created 11/05/2021)
- 📄 [Amendment 55694.A1](#) (Created 03/06/2020)
- 📄 [Original Submission 55694](#) (Created 03/03/2020)

User Uploaded Documents i

Title	Version Number	Document Date	Document Type	Size
1. <span style="font-size: 0.8em;">📄</span> <a href="#">55694.A2_Ammendment-55694-A1_PEERWISE+QMP--Ethics Application Form for Secondary Data Analysis</a> <span style="font-size: 0.8em;">📄</span>	3.0	11/05/2021	Ethics Form	1,264 Kb

Checklist

Submission Questionnaire ✔

Attachments ✔

Coordinators

Adriana Wilde (agw106@ecs.soton.ac.uk)

David Millard (dem@soton.ac.uk) supervisor

→ Create Amendment
🚫 Abandon Study

## Ethics Application Form for SECONDARY DATA ANALYSIS

Version September 2019

*Please consult the guidance at the end of this form before completing and submitting your application.*

1. **Name(s):** Adriana Wilde
2. **Current Position:** PhD researcher
3. **Contact Details:**  
**Division:** FEPS / ECS (WAIS)  
**Email:** agw106@soton.ac.uk  
**Phone:** 023 8059 9039
4. **Is your research being conducted as part of an education qualification?**  
Yes  No
5. **If Yes, please give the name of your supervisor:**  
David Millard (dem@ecs.soton.ac.uk)
6. **Title of your research project / study:**  
Learner Engagement within Peer-Supported Environments mediated by Digital Technologies: PeerWise and MOOC Platforms
7. **Briefly describe the rationale, aims, design and research questions of your research**

*Please indicate clearly whether you are applying for ethics approval for a specific piece of research, or for overarching ethics approval to use certain datasets for a range of research activities. Approval for the latter will only cover the datasets specified here, for a maximum of 3 years and then subject to renewal.*

The rationale of this research lies on the need to understand the patterns of learner engagement using digital technologies in peer-supported environments, such as with *PeerWise* in the context of face-to-face instruction, and also in the context of massive open online courses (MOOCs).

I aim to identify the main patterns of engagement using a machine learning technique that involves the grouping of individual data points (*clustering*). In addition to the engagement clusters, the data will be used to predict attainment. In more general terms, the research aims to identify whether a specific pattern of engagement is more indicative than others to achieve academic success, i.e. higher marks in the case of face-to-face instruction or, in the case of MOOCs, retention and completion.

This research rest upon observations made whilst lecturing a module in Computer Science, where I was responsible for the assessment design and learning activities of two consecutive cohorts in this particular module. With others in my teaching team, I observed that the *PeerWise* tool had been used as a revision aid beyond the requirements of the module. The use of the tool and these observations have been reported previously

as listed below [1,2,3,4]. It also builds upon our prior research on learning analytics on MOOCs [5,6,7,8], and in our hypothesis that there are significant commonalities to be drawn between the way students behaviour in face-to-face environments and in MOOCs [9] and how these behaviours impact on their academic success.

Therefore, I would like overarching ethics approval to use the datasets regarding students' participation through the *PeerWise* software which was used for students' authoring and answering Multiple-Choice Questions (MCQ) on the 2016/17 and 2015/16 cohorts of the COMP2213 module (see section 8 for details). In addition, I would like to use the data of the students' attainment, both in the *QuestionMark Perception* exam (which had a 50% element of MCQ) and the overall marks in the module (which include the coursework element), based on preliminary analysis of the cohort data as done to inform my teaching practice whilst a lecturer in the module during the periods under consideration.

The aim is to investigate whether the learner engagement within *PeerWise* has had a positive effect on their learning as reflected on the summative assessment within the module. The hypothesis being that those students who were highly engaged with the tool, performed better in the formal assessment.

The design includes the creation of a feature vector associated to each student (which will be anonymised at source), to characterise their "engagement and learning profiles". I would also like to identify those features which are most highly predictive of a good performance in the exam.

This design is mirrored in a study of learner engagement in the "*Understanding language*" FutureLearn MOOC for the first six runs of this course (November 2014 - May2017). In this case, a feature vector is constructed to characterise learner engagement, and identify those features which are predictive of retention and completion in the course.

The research question is whether we can see the same learner behaviour manifesting in different ways in these two peer-supported environments or whether different behaviours altogether are presented.

- [1] A. Wilde (2019) [Rising to Challenges in Assessment and Feedback in HCI Education: A-Peer-Supported Approach](http://edshare.soton.ac.uk/20154/) (<http://edshare.soton.ac.uk/20154/>)
- [2] S. Snow, A. Wilde, m.c. schraefel, and P. Denny (2019) "A discursive question: Supporting student-authored multiple-choice questions through peer-learning software in non-STEMM disciplines". *British Journal of Educational Technology* 50 (4), 1815-1830. (<https://doi.org/10.1111/bjet.12686>).
- [3] A. Wilde and S. Snow (2018) "Addressing challenges in assessing Human-Computer Interaction at scale". In the Computing Education Practice conference, 11-12 January, Durham, UK.
- [4] S. Snow and A. Wilde (2017) "Supporting Authoring of Multiple-Choice Questions in Human-Computer Interaction using PeerWise". In the conference *What Works in Assessment and Feedback: Simply Better*, 14 September, Southampton, UK.
- [5] A. Wilde, M. León, and K. Borthwick (2017) "Understanding Language: Understanding MOOC learners" *Rosell-Aguilar, Fernando, Beaven, Tita and Fuertes-Gutierrez, Mara (eds.) In Proceedings of the 7th Annual Conference in the Innovative Language Teaching and Learning at University.*
- [6] A. Sunar, G. Dogan, A. Wilde, and I. Duru (2017) "Leveraging learning analytics to identify and overcome barriers to MOOCs in a foreign language" *EMOOCs 2017, Madrid, Spain. 22-26 May.*

- [7] R. Cobos, A. Wilde, and E. Zaluska (2017) "Predicting attrition from massive open online courses in FutureLearn and edX" *FutureLearn data: what we currently have, what we are learning and how it is demonstrating Learning in MOOCs. Workshop at the 7th International Learning Analytics and Knowledge Conference. Simon Fraser University, Vancouver, Canada, 13-17 March, p. 74-93*
- [8] A. Wilde (2016) "Understanding persuasive technologies to improve completion rates in MOOCs" *HCI and the Educational Technology Revolution. Workshop at the International Conference on Advanced Visual Interfaces (AVI 2016), Bari, Italy, 7 June.*
- [9] A. Wilde, E. Zaluska, and D. Millard (2015) "Student success on face-to-face instruction and MOOCs." *Web Science Education: Curriculum, MOOCs and Learning. WEB SCIENCE 2015.*

## 8. Describe the data you wish to analyse

*Please give details of the title of the dataset, nature of data subjects (e.g. individuals or organisations), thematic focus and country/countries covered. Indicate whether the data are qualitative or quantitative, survey data, administrative data or other types of data. Identify the source from where you will be obtaining the data (including a web address where appropriate).*

As the title of the research suggests, there are two distinct parts of the study which use a similar approach to data processing and analysis in two different peer-supported environments.

For the first part of the study (engagement in *PeerWise*), the data subjects are students of the second-year module Interaction Design (COMP2213), a compulsory module for Computer Science at the University of Southampton. Specifically, the data subjects belong the cohorts of 2015/16 and 2016/17, as I was primarily responsible for their formal assessment as member of the teaching team. These subjects were asked at the time to use the *PeerWise* software (<https://peerwise.cs.auckland.ac.nz>) as part of the module coursework to create multiple-choice questions on the module topics, which in turn were answered (and commented upon) by their peers.

There are two data sources for this part of the study: firstly, students' participation in the module via the free software **PeerWise**, which on registration students agreed that could be used for research purposes (previously approved as ERGO/FPSE/20318). Secondly, their attainment data which have been used to evaluate their learning within the module via the university-managed software **QuestionMark Perception**. The data in *PeerWise* are predominantly quantitative, reflecting their engagement with the module by their timestamped activity (e.g. creation of multiple-choice questions, provision of answers, ratings given and received on created questions, comments, number of replies given, number of followers, badges obtained, and so on). However, there are also some qualitative data, such as the text to the actual questions, and comments. The assessment data consist on the answers provided in the *QuestionMark Perception* exam, both in the MCQ element of the exam and in the free text. The marks obtained in the coursework element of the assessment will also be used. As explained above, these two datasets will be combined with the purpose of creating a feature vector to characterise each individual student engagement and attainment.

For the second part of the study, engagement data of participants in the "*Understanding language*" FutureLearn MOOC for the first six runs of this course

(November 2014 – May 2017). For this part of the study, the engagement data available across various files allows for a learner profile which includes, for each week of the course, the number of articles read, number of videos viewed, discussions joined, number of completed assignments, generated comments and likes received.

**9. What are the terms and conditions around the use of the data? Did data subjects give consent for their data to be re-used? If not, on what basis is re-use of the data justified?**

*Please state what (if any) conditions the data archive imposes (e.g. registration, signing of confidentiality agreement, specific training etc.). In many cases the data controller will have given explicit permission for data re-use. Please explain how you justify the use of data if approval and consents for the original data collection and re-use are not in place. This may be the case where, for example, the original data collection predated requirements for ethics review or occurred in a jurisdiction where explicit consent and approval are not required.*

Yes, on registration, data subjects were told the repository could be used for research in Computer Science Education. This consent was gathered under the study “Writing academic papers based on de-identified data taken from students’ participation in and reflections upon *PeerWise* software used in COMP2213 (ERGO FPSE/20318). Similarly, *FutureLearn* gathered such consent for all of the six runs of the course *Understanding Language*, between November 2014 and May 2017, to be used in this study.

**10. Do you intend to process personal data (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data>) that are sensitive (‘special category’) personal data as defined by the the Data Protection Act 2018 following the General Data Protection Regulation (GDPR) (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/>), or data relating to a person’s criminal convictions, even if such data are publicly available and/or have been pseudonymised (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/>)?**

Yes  No

If YES, please specify what personal data will be processed and why.

The student ID of each student is the unique identifier which links all the different files in the datasets. It is required to create the feature vector above described, however it will be replaced with another unique identifier which cannot be traced back to any individual.

**11. Do you intend to link two or more datasets?**

*Data linkage refers to merging of information from two or more sources of data to consolidate facts concerning an individual or an event that are not available in any separate record. Please note that for the purposes of research ethics we are*

*not interested in the merging of different waves of a particular survey, or the merging of data from different countries for the same survey.*

Yes  No

If YES, please give details of which datasets will be linked and for what purposes.

See answers to question 7 and 8.

**12. How will you store and manage the data before and during the analysis? What will happen with the data at the end of the project?**

*Please consult the University of Southampton's Research Data Management Policy (<http://library.soton.ac.uk/researchdata/storage> and <http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html>), and indicate how you will abide by it.*

The data will be stored in a password-protected university-machine during the pre-processing phase of the study. Copies of the de-identified data will be stored in the University's network storage during the analysis phase, and will be deleted at the end of the research project.

**13. How will you minimise the risk that data subjects (individuals or organisations) could be identified in your presentation of results?**

*Please consider whether disclosive ID codes have been used (e.g. date of birth) and whether it is theoretically possible to identify individuals by combining characteristics (e.g. widow in Hampshire with 14 children) or by combining datasets. How will you protect individuals' anonymity in your analysis and dissemination?*

It is not possible to identify individuals through the combination of the datasets, as the only personal information used is the student ID, which is used for the purpose of linking the two datasets (for PeerWise and QMP) but it is replaced with another unique identifier when pre-processing the data. This effectively means that the link to the original identity of each individual will be destroyed before the data analysis stage.

**14. What other ethical risks are raised by your research, and how do you intend to manage these?**

*Issues may arise due to the nature of the research you intend to undertake and/or the subject matter of the data. Examples include: data or analysis that are culturally or socially sensitive; data relating to criminal activity, including terrorism, and security sensitive issues.*

none

**15. Please outline any other information that you feel may be relevant to this submission.**

*For example, will you be using the services or facilities of ONS, ADRN, or HSCIC and/or are you obtaining ethical review from NRES (through IRAS) or other? Please confirm whether the data being used are already in the public domain.*

None other than those explained above.

16. **Please indicate if you, your supervisor or a member of the study team/research group (including any institution that they act for, if different from the University) are a data controller and/or data processor in relation to the personal data you intend to process as defined by the Data Protection Act 2018 following the GDPR, and confirm that you/they understand your/their respective responsibilities ( <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/controllers-and-processors//>)**

I am both a data controller and data processor, as follows:

- For the *PeerWise* data: I have been a jointly data controller with Paul Denny, from the University of Auckland in New Zealand. Paul is the creator of PeerWise and as such, he holds the dataset associated with the data-subjects participation in the COMP2213 course within *PeerWise*.
- For the *QuestionMark Perception* data: I have been a jointly data controller with Steve Snow (ss33g15, for both cohorts under consideration), m.c. schraefel (mc, for the 2015/16 cohort only), and Nick Gibbins (nmg, for the 2016/17 cohort only), as we were all examiners and had access to the exam files.

I am the only data processor with access to both sets of data and I fully understand my responsibility to de-anonymise the combined data prior the creation of the feature vector as explained in point 7. Similarly, for the FutureLearn dataset, I do not use any data which could lead to the identification of any individuals.



## Guidance on applying for ethics approval for secondary data analysis

If your research PURELY involves the following, you do not need to apply for ethics approval:

- analysis of aggregated data on individuals or organisations (e.g. GDP, labour force participation rates, fertility rates);
- meta-analyses (i.e. the analysis of studies);
- literature reviews or reviews/analyses of reports, policies, documents, meeting minutes, newspaper articles, films.

### Filling in the online submission form on ERGO II:

- Please give your application a title that includes 'SDA' (Secondary Data Analysis).
- Please refer to the "**ERGO II Guidance for Applicants**" document (downloadable from the ERGO II site) on how to answer the submission questionnaire correctly..

### Additional Forms:

If your study PURELY involves secondary analysis of data, you only need to fill in the 'Ethics Application Form for Secondary Data Analysis'. You do not need a Risk Assessment Form.

If your study is a mixed-method study involving secondary data analysis AND some component of data collection (e.g. interviews, online survey), then you need to fill in additional forms:

- Ethics Application Form (for studies other than secondary data analysis)
- Risk Assessment Form
- Participant Information Sheet
- Consent Form
- Draft research instrument

### Please note:

- You must not begin data analysis until ethical approval has been obtained.
- It is your responsibility to follow the University of Southampton's Ethics Policy and any relevant academic or professional guidelines in the conduct of your research. This includes ensuring confidentiality in the storage and use of data.
- It is your responsibility to provide full and accurate information in completing this form.

## Ethics Application Form for SECONDARY DATA ANALYSIS

Version September 2019

*Please consult the guidance at the end of this form before completing and submitting your application.*

1. **Name(s):** Adriana Wilde
2. **Current Position:** PhD researcher
3. **Contact Details:**  
**Division:** FEPS / ECS (WAIS)  
**Email:** agw106@soton.ac.uk  
**Phone:** 023 8059 9039

4. **Is your research being conducted as part of an education qualification?**  
Yes  No

5. **If Yes, please give the name of your supervisor:**

David Millard (dem@ecs.soton.ac.uk)

6. **Title of your research project / study:**

Learner Engagement within Peer-Supported Environments mediated by Digital Technologies: PeerWise and MOOC Platforms

7. **Briefly describe the rationale, aims, design and research questions of your research**

*Please indicate clearly whether you are applying for ethics approval for a specific piece of research, or for overarching ethics approval to use certain datasets for a range of research activities. Approval for the latter will only cover the datasets specified here, for a maximum of 3 years and then subject to renewal.*

The rationale of this research lies on the need to understand the patterns of learner engagement using digital technologies in peer-supported environments, such as with *PeerWise* in the context of face-to-face instruction, and also in the context of massive open online courses (MOOCs).

I aim to identify the main patterns of engagement using a machine learning technique that involves the grouping of individual data points (*clustering*). In addition to the engagement clusters, the data will be used to predict attainment. In more general terms, the research aims to identify whether a specific pattern of engagement is more indicative than others to achieve academic success, i.e. higher marks in the case of face-to-face instruction or, in the case of MOOCs, retention and completion.

This research rest upon observations made whilst lecturing a module in Computer Science, where I was responsible for the assessment design and learning activities of two consecutive cohorts in this particular module. With others in my teaching team, I observed that the *PeerWise* tool had been used as a revision aid beyond the requirements of the module. The use of the tool and these observations have been reported previously

as listed below [1,2,3,4]. It also builds upon our prior research on learning analytics on MOOCs [5,6,7,8], and in our hypothesis that there are significant commonalities to be drawn between the way students behaviour in face-to-face environments and in MOOCs [9] and how these behaviours impact on their academic success.

Therefore, I would like overarching ethics approval to use the datasets regarding students' participation through the *PeerWise* software which was used for students' authoring and answering Multiple-Choice Questions (MCQ) on the 2016/17 and 2015/16 cohorts of the COMP2213 module (see section 8 for details). In addition, I would like to use the data of the students' attainment, both in the *QuestionMark Perception* exam (which had a 50% element of MCQ) and the overall marks in the module (which include the coursework element), based on preliminary analysis of the cohort data as done to inform my teaching practice whilst a lecturer in the module during the periods under consideration.

The aim is to investigate whether the learner engagement within *PeerWise* has had a positive effect on their learning as reflected on the summative assessment within the module. The hypothesis being that those students who were highly engaged with the tool, performed better in the formal assessment.

The design includes the creation of a feature vector associated to each student (which will be anonymised at source), to characterise their "engagement and learning profiles". I would also like to identify those features which are most highly predictive of a good performance in the exam.

This design is mirrored in a study of learner engagement in the "*Understanding language*" FutureLearn MOOC for all eleven runs of this course, as well as in all of the six run of the "Archaeology of Portus: Exploring the Lost Harbour of Ancient Rome" FutureLearn MOOC (Portus). In these cases, a feature vector is constructed to characterise learner engagement, and identify those features which are predictive of retention and completion in the course.

The research question is whether we can see the same learner behaviour manifesting in different ways in these two peer-supported environments or whether different behaviours altogether are presented.

- [1] A. Wilde (2019) [Rising to Challenges in Assessment and Feedback in HCI Education: A-Peer-Supported Approach](http://edshare.soton.ac.uk/20154/) (<http://edshare.soton.ac.uk/20154/>)
- [2] S. Snow, A. Wilde, m.c. schraefel, and P. Denny (2019) "A discursive question: Supporting student-authored multiple-choice questions through peer-learning software in non-STEMM disciplines". *British Journal of Educational Technology* 50 (4), 1815-1830. (<https://doi.org/10.1111/bjet.12686>).
- [3] A. Wilde and S. Snow (2018) "Addressing challenges in assessing Human-Computer Interaction at scale". In the Computing Education Practice conference, 11-12 January, Durham, UK.
- [4] S. Snow and A. Wilde (2017) "Supporting Authoring of Multiple-Choice Questions in Human-Computer Interaction using PeerWise". In the conference *What Works in Assessment and Feedback: Simply Better*, 14 September, Southampton, UK.
- [5] A. Wilde, M. León, and K. Borthwick (2017) "Understanding Language: Understanding MOOC learners" *Rosell-Aguilar, Fernando, Beaven, Tita and Fuentes-Gutierrez, Mara (eds.) In Proceedings of the 7th Annual Conference in the Innovative Language Teaching and Learning at University.*

- [6] A. Sunar, G. Dogan, A. Wilde, and I. Duru (2017) "Leveraging learning analytics to identify and overcome barriers to MOOCs in a foreign language" *EMOOCs 2017, Madrid, Spain. 22-26 May.*
- [7] R. Cobos, A. Wilde, and E. Zaluska (2017) "Predicting attrition from massive open online courses in FutureLearn and edX" *FutureLearn data: what we currently have, what we are learning and how it is demonstrating learning in MOOCs. Workshop at the 7th International Learning Analytics and Knowledge Conference. Simon Fraser University, Vancouver, Canada, 13-17 March, p. 74-93*
- [8] A. Wilde (2016) "Understanding persuasive technologies to improve completion rates in MOOCs" *HCI and the Educational Technology Revolution. Workshop at the International Conference on Advanced Visual Interfaces (AVI 2016), Bari, Italy, 7 June.*
- [9] A. Wilde, E. Zaluska, and D. Millard (2015) "Student success on face-to-face instruction and MOOCs." *Web Science Education: Curriculum, MOOCs and Learning. WEB SCIENCE 2015.*

## 8. Describe the data you wish to analyse

*Please give details of the title of the dataset, nature of data subjects (e.g. individuals or organisations), thematic focus and country/countries covered. Indicate whether the data are qualitative or quantitative, survey data, administrative data or other types of data. Identify the source from where you will be obtaining the data (including a web address where appropriate).*

As the title of the research suggests, there are two distinct parts of the study which use a similar approach to data processing and analysis in two different peer-supported environments.

For the first part of the study (engagement in *PeerWise*), the data subjects are students of the second-year module Interaction Design (COMP2213), a compulsory module for Computer Science at the University of Southampton. Specifically, the data subjects belong the cohorts of 2015/16 and 2016/17, as I was primarily responsible for their formal assessment as member of the teaching team. These subjects were asked at the time to use the *PeerWise* software (<https://peerwise.cs.auckland.ac.nz>) as part of the module coursework to create multiple-choice questions on the module topics, which in turn were answered (and commented upon) by their peers.

There are two data sources for this part of the study: firstly, students' participation in the module via the free software ***PeerWise***, which on registration students agreed that could be used for research purposes (previously approved as ERGO/FPSE/20318). Secondly, their attainment data which have been used to evaluate their learning within the module via the university-managed software ***QuestionMark Perception***. The data in *PeerWise* are predominantly quantitative, reflecting their engagement with the module by their timestamped activity (e.g. creation of multiple-choice questions, provision of answers, ratings given and received on created questions, comments, number of replies given, number of followers, badges obtained, and so on). However, there are also some qualitative data, such as the text to the actual questions, and comments. The assessment data consist on the answers provided in the *QuestionMark Perception* exam, both in the MCQ element of the exam and in the free text. The marks obtained in the coursework element of the assessment will also be used. As explained above, these two datasets will be combined with the purpose of creating a feature vector to characterise each individual student engagement and attainment.

For the second part of the study, engagement data of participants in the “*Understanding language*” FutureLearn MOOC for all of the eleven runs of this course, as well as for all of the six runs of the “Portus” course. For this part of the study, the engagement data available across various files allows for a learner profile which includes, for each week of the course, the number of articles read, number of videos viewed, discussions joined, number of completed assignments, generated comments and likes received. Other files of this dataset include survey responses within the platform which are also valuable to judge the learners’ engagement and attainment.

**9. What are the terms and conditions around the use of the data? Did data subjects give consent for their data to be re-used? If not, on what basis is re-use of the data justified?**

*Please state what (if any) conditions the data archive imposes (e.g. registration, signing of confidentiality agreement, specific training etc.). In many cases the data controller will have given explicit permission for data re-use. Please explain how you justify the use of data if approval and consents for the original data collection and re-use are not in place. This may be the case where, for example, the original data collection predated requirements for ethics review or occurred in a jurisdiction where explicit consent and approval are not required.*

Yes, on registration, data subjects were told the repository could be used for research in Computer Science Education. This consent was gathered under the study “Writing academic papers based on de-identified data taken from students’ participation in and reflections upon *PeerWise* software used in COMP2213 (ERGO FPSE/20318). Similarly, *FutureLearn* gathered such consent for all of the runs of all of the courses to be used in this study.

**10. Do you intend to process personal data (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data>) that are sensitive (‘special category’) personal data as defined by the the Data Protection Act 2018 following the General Data Protection Regulation (GDPR) (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/>), or data relating to a person’s criminal convictions, even if such data are publicly available and/or have been pseudonymised (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/>)?**

Yes  No

If YES, please specify what personal data will be processed and why.

The student ID of each student is the unique identifier which links all the different files in the datasets. It is required to create the feature vector above described, however, even though it will be replaced with another unique identifier, there is a possibility that some learners may have disclosed ‘special category’ personal data within their comments (especially during the introductory week in both courses), which could lead to the identification of these individuals.

**11. Do you intend to link two or more datasets?**

*Data linkage refers to merging of information from two or more sources of data to consolidate facts concerning an individual or an event that are not available in any separate record. Please note that for the purposes of research ethics we are not interested in the merging of different waves of a particular survey, or the merging of data from different countries for the same survey.*

Yes  No

If YES, please give details of which datasets will be linked and for what purposes.

See answers to question 7 and 8.

**12. How will you store and manage the data before and during the analysis? What will happen with the data at the end of the project?**

*Please consult the University of Southampton's Research Data Management Policy (<http://library.soton.ac.uk/researchdata/storage> and <http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html>), and indicate how you will abide by it.*

The data will be stored in a password-protected university-machine during the pre-processing phase of the study. Copies of the de-identified data will be stored in the University's network storage during the analysis phase, and will be deleted at the end of the research project.

**13. How will you minimise the risk that data subjects (individuals or organisations) could be identified in your presentation of results?**

*Please consider whether disclosive ID codes have been used (e.g. date of birth) and whether it is theoretically possible to identify individuals by combining characteristics (e.g. widow in Hampshire with 14 children) or by combining datasets. How will you protect individuals' anonymity in your analysis and dissemination?*

As mentioned in the amended part of question 10, it may be possible to identify individuals through the combination of the datasets (e.g. comments files, survey responses). However, I will minimise the risk by pre-processing the free-text in the comments files and extracting numerical features indicative of the learning engagement. After this, I will delete the free text from the dataset.

With regards to the survey responses, I will only filter content related to course engagement (and disregard any other that may make reference to personal characteristics). In reporting my findings, I would report group characteristics (rather than individuals), and, if appropriate, quote any survey responses after stripping the anonymised userID of the learner assigned by FutureLearn, and assigning it a new one (e.g. "Learner A").

The above mentioned risk does not exist with the part of the study involving students in face-to-face instruction, as the only personal information used is the student ID, which is used for the purpose of linking the two datasets (for PeerWise and QMP) but it is replaced with another unique identifier when pre-processing the data. This effectively

Delete

Delete

Format

means that the link to the original identity of each individual will be destroyed before the data analysis stage.

**14. What other ethical risks are raised by your research, and how do you intend to manage these?**

*Issues may arise due to the nature of the research you intend to undertake and/or the subject matter of the data. Examples include: data or analysis that are culturally or socially sensitive; data relating to criminal activity, including terrorism, and security sensitive issues.*

none

**15. Please outline any other information that you feel may be relevant to this submission.**

*For example, will you be using the services or facilities of ONS, ADRN, or HSCIC and/or are you obtaining ethical review from NRES (through IRAS) or other? Please confirm whether the data being used are already in the public domain.*

None other than those explained above.

**16. Please indicate if you, your supervisor or a member of the study team/research group (including any institution that they act for, if different from the University) are a data controller and/or data processor in relation to the personal data you intend to process as defined by the Data Protection Act 2018 following the GDPR, and confirm that you/they understand your/their respective responsibilities (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/controllers-and-processors/>)**

I am both a data controller and data processor, as follows:

- For the PeerWise data: I have been a jointly data controller with Paul Denny, from the University of Auckland in New Zealand. Paul is the creator of PeerWise and as such, he holds the dataset associated with the data-subjects participation in the COMP2213 course within PeerWise.
- For the QuestionMark Perception data: I have been a jointly data controller with Steve Snow (ss33g15, for both cohorts under consideration), m.c. schraefel (mc, for the 2015/16 cohort only), and Nick Gibbins (nmg, for the 2016/17 cohort only), as we were all examiners and had access to the exam files.

I am the only data processor with access to both sets of data and I fully understand my responsibility to de-anonymise the combined data prior the creation of the feature vector (and qualitative analysis of the survey data) as explained in points 7, 10 and 13.

Dele  
do r  
iden



## Guidance on applying for ethics approval for secondary data analysis

If your research PURELY involves the following, you do not need to apply for ethics approval:

- analysis of aggregated data on individuals or organisations (e.g. GDP, labour force participation rates, fertility rates);
- meta-analyses (i.e. the analysis of studies);
- literature reviews or reviews/analyses of reports, policies, documents, meeting minutes, newspaper articles, films.

### Filling in the online submission form on ERGO II:

- Please give your application a title that includes 'SDA' (Secondary Data Analysis).
- Please refer to the "**ERGO II Guidance for Applicants**" document (downloadable from the ERGO II site) on how to answer the submission questionnaire correctly..

### Additional Forms:

If your study PURELY involves secondary analysis of data, you only need to fill in the 'Ethics Application Form for Secondary Data Analysis'. You do not need a Risk Assessment Form.

If your study is a mixed-method study involving secondary data analysis AND some component of data collection (e.g. interviews, online survey), then you need to fill in additional forms:

- Ethics Application Form (for studies other than secondary data analysis)
- Risk Assessment Form
- Participant Information Sheet
- Consent Form
- Draft research instrument

### Please note:

- You must not begin data analysis until ethical approval has been obtained.
- It is your responsibility to follow the University of Southampton's Ethics Policy and any relevant academic or professional guidelines in the conduct of your research. This includes ensuring confidentiality in the storage and use of data.
- It is your responsibility to provide full and accurate information in completing this form.



## Ethics Application Form for SECONDARY DATA ANALYSIS

*Version September 2019*

***Please consult the guidance at the end of this form before completing and submitting your application.***

1. **Name(s):** Adriana Wilde
2. **Current Position:** PhD researcher
3. **Contact Details:**  
**Division:** FEPS / ECS (WAIS)  
**Email:** agw106@soton.ac.uk  
**Phone:** 023 8059 9039
4. **Is your research being conducted as part of an education qualification?**  
**Yes**  **No**
5. **If Yes, please give the name of your supervisor:**  
 David Millard (dem@ecs.soton.ac.uk)
6. **Title of your research project / study:**  
 Learner Engagement within Peer-Supported Environments mediated by Digital Technologies: PeerWise and MOOC Platforms
7. **Briefly describe the rationale, aims, design and research questions of your research**

*Please indicate clearly whether you are applying for ethics approval for a specific piece of research, or for overarching ethics approval to use certain datasets for a range of research activities. Approval for the latter will only cover the datasets specified here, for a maximum of 3 years and then subject to renewal.*

The rationale of this research lies on the need to understand the patterns of learner engagement using digital technologies in peer-supported environments, such as with *PeerWise* in the context of face-to-face instruction, and also in the context of massive open online courses (MOOCs).

I aim to identify the main patterns of engagement using a machine learning technique that involves the grouping of individual data points (*clustering*). In addition to the engagement clusters, the data will be used to predict attainment. In more general terms, the research aims to identify whether a specific pattern of engagement is more indicative than others to achieve academic success, i.e. higher marks in the case of face-to-face instruction or, in the case of MOOCs, retention and completion.

This research rest upon observations made whilst lecturing a module in Computer Science, where I was responsible for the assessment design and learning activities of two consecutive cohorts in this particular module. With others in my teaching team, I observed that the *PeerWise* tool had been used as a revision aid beyond the requirements of the module. The use of the tool and these observations have been reported previously as listed below [1,2,3,4]. It also builds upon our prior research on learning analytics on MOOCs [5,6,7,8], and in our hypothesis that there are significant commonalities to be

drawn between the way students behaviour in face-to-face environments and in MOOCs [9] and how these behaviours impact on their academic success.

Therefore, I would like overarching ethics approval to use the datasets regarding students' participation through the *PeerWise* software which was used for students' authoring and answering Multiple-Choice Questions (MCQ) on the 2016/17 and 2015/16 cohorts of the COMP2213 module (see section 8 for details). In addition, I would like to use the data of the students' attainment, both in the *QuestionMark Perception* exam (which had a 50% element of MCQ) and the overall marks in the module (which include the coursework element), based on preliminary analysis of the cohort data as done to inform my teaching practice whilst a lecturer in the module during the periods under consideration.

The aim is to investigate whether the learner engagement within *PeerWise* has had a positive effect on their learning as reflected on the summative assessment within the module. The hypothesis being that those students who were highly engaged with the tool, performed better in the formal assessment.

The design includes the creation of a feature vector associated to each student (which will be anonymised at source), to characterise their "engagement and learning profiles". I would also like to identify those features which are most highly predictive of a good performance in the exam.

This design is mirrored in a study of learner engagement in the "*Understanding language*" FutureLearn MOOC for all eleven runs of this course, as well as in all of the six run of the "*Archaeology of Portus: Exploring the Lost Harbour of Ancient Rome*" FutureLearn MOOC (*Portus*). In these cases, a feature vector is constructed to characterise learner engagement, and identify those features which are predictive of retention and completion in the course.

The research question is whether we can see the same learner behaviour manifesting in different ways in these two peer-supported environments or whether different behaviours altogether are presented.

- [1] A. Wilde (2019) [Rising to Challenges in Assessment and Feedback in HCI Education: A-Peer-Supported Approach](http://edshare.soton.ac.uk/20154/) (<http://edshare.soton.ac.uk/20154/>)
- [2] S. Snow, A. Wilde, m.c. schraefel, and P. Denny (2019) "A discursive question: Supporting student-authored multiple-choice questions through peer-learning software in non-STEMM disciplines". *British Journal of Educational Technology* 50 (4), 1815-1830. (<https://doi.org/10.1111/bjet.12686>).
- [3] A. Wilde and S. Snow (2018) "Addressing challenges in assessing Human-Computer Interaction at scale". In the Computing Education Practice conference, 11-12 January, Durham, UK.
- [4] S. Snow and A. Wilde (2017) "Supporting Authoring of Multiple-Choice Questions in Human-Computer Interaction using PeerWise". In the conference *What Works in Assessment and Feedback: Simply Better*, 14 September, Southampton, UK.
- [5] A. Wilde, M. León, and K. Borthwick (2017) "Understanding Language: Understanding MOOC learners" *Rosell-Aguilar, Fernando, Beaven, Tita and Fuertes-Gutierrez, Mara (eds.) In Proceedings of the 7th Annual Conference in the Innovative Language Teaching and Learning at University.*
- [6] A. Sunar, G. Dogan, A. Wilde, and I. Duru (2017) "Leveraging learning analytics to identify and overcome barriers to MOOCs in a foreign language" *EMOOCs 2017, Madrid, Spain. 22-26 May.*
- [7] R. Cobos, A. Wilde, and E. Zaluska (2017) "Predicting attrition from massive open online courses in FutureLearn and edX" *FutureLearn data: what we currently have, what we are learning and how it is demonstrating*

- Learning in MOOCs. Workshop at the 7th International Learning Analytics and Knowledge Conference. Simon Fraser University, Vancouver, Canada, 13-17 March, p. 74-93*
- [8] A. Wilde (2016) "Understanding persuasive technologies to improve completion rates in MOOCs" *HCI and the Educational Technology Revolution. Workshop at the International Conference on Advanced Visual Interfaces (AVI 2016), Bari, Italy, 7 June.*
- [9] A. Wilde, E. Zaluska, and D. Millard (2015) "Student success on face-to-face instruction and MOOCs." *Web Science Education: Curriculum, MOOCs and Learning. WEB SCIENCE 2015.*

## 8. Describe the data you wish to analyse

*Please give details of the title of the dataset, nature of data subjects (e.g. individuals or organisations), thematic focus and country/countries covered. Indicate whether the data are qualitative or quantitative, survey data, administrative data or other types of data. Identify the source from where you will be obtaining the data (including a web address where appropriate).*

As the title of the research suggests, there are two distinct parts of the study which use a similar approach to data processing and analysis in two different peer-supported environments.

For the first part of the study (engagement in *PeerWise*), the data subjects are students of the second-year module Interaction Design (COMP2213), a compulsory module for Computer Science at the University of Southampton. Specifically, the data subjects belong to the cohorts of 2015/16 and 2016/17, as I was primarily responsible for their formal assessment as member of the teaching team. These subjects were asked at the time to use the *PeerWise* software (<https://peerwise.cs.auckland.ac.nz>) as part of the module coursework to create multiple-choice questions on the module topics, which in turn were answered (and commented upon) by their peers.

There are two data sources for this part of the study: firstly, students' participation in the module via the free software *PeerWise*, which on registration students agreed that could be used for research purposes (previously approved as ERGO/FPSE/20318). Secondly, their attainment data which have been used to evaluate their learning within the module via the university-managed software *QuestionMark Perception*. The data in *PeerWise* are predominantly quantitative, reflecting their engagement with the module by their timestamped activity (e.g. creation of multiple-choice questions, provision of answers, ratings given and received on created questions, comments, number of replies given, number of followers, badges obtained, and so on). However, there are also some qualitative data, such as the text to the actual questions, and comments. The assessment data consist of the answers provided in the *QuestionMark Perception* exam, both in the MCQ element of the exam and in the free text. The marks obtained in the coursework element of the assessment will also be used. As explained above, these two datasets will be combined with the purpose of creating a feature vector to characterise each individual student engagement and attainment.

For the second part of the study, engagement data of participants in the "*Understanding language*" FutureLearn MOOC for all of the eleven runs of this course, as well as for all of the six runs of the "*Portus*" course. For this part of the study, the engagement data available across various files allows for a learner profile which includes, for each week of the course, the number of articles read, number of videos viewed, discussions joined, number of completed assignments, generated comments and likes

received. Other files of this dataset include survey responses within the platform which are also valuable to judge the learners' engagement and attainment.

**9. What are the terms and conditions around the use of the data? Did data subjects give consent for their data to be re-used? If not, on what basis is re-use of the data justified?**

*Please state what (if any) conditions the data archive imposes (e.g. registration, signing of confidentiality agreement, specific training etc.). In many cases the data controller will have given explicit permission for data re-use. Please explain how you justify the use of data if approval and consents for the original data collection and re-use are not in place. This may be the case where, for example, the original data collection predated requirements for ethics review or occurred in a jurisdiction where explicit consent and approval are not required.*

Yes, on registration, data subjects were told the repository could be used for research in Computer Science Education. This consent was gathered under the study "Writing academic papers based on de-identified data taken from students' participation in and reflections upon *PeerWise* software used in COMP2213 (ERGO FPSE/20318). Similarly, *FutureLearn* gathered such consent for all of the runs of all of the courses to be used in this study.

**10. Do you intend to process personal data (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data>) that are sensitive ('special category') personal data as defined by the the Data Protection Act 2018 following the General Data Protection Regulation (GDPR) (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/>), or data relating to a person's criminal convictions, even if such data are publicly available and/or have been pseudonymised (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/>)?**

Yes  No

If YES, please specify what personal data will be processed and why.

The student ID of each student is the unique identifier which links all the different files in the datasets. It is required to create the feature vector above described, however, even though it will be replaced with another unique identifier, there is a possibility that some learners may have disclosed 'special category' personal data within their comments (especially during the introductory week in both courses), which could lead to the identification of these individuals.

**11. Do you intend to link two or more datasets?**

*Data linkage refers to merging of information from two or more sources of data to consolidate facts concerning an individual or an event that are not available in any separate record. Please note that for the purposes of research ethics we are*

*not interested in the merging of different waves of a particular survey, or the merging of data from different countries for the same survey.*

Yes  No

If YES, please give details of which datasets will be linked and for what purposes.

See answers to question 7 and 8.

**12. How will you store and manage the data before and during the analysis?  
What will happen with the data at the end of the project?**

*Please consult the University of Southampton's Research Data Management Policy (<http://library.soton.ac.uk/researchdata/storage> and <http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html>), and indicate how you will abide by it.*

The data will be stored in a password-protected university-machine during the pre-processing phase of the study. Copies of the de-identified data will be stored in the University's network storage during the analysis phase, and will be further processed to protect the rights of the data subjects, honouring the original consent as detailed in question 9, above.

In particular, a clean dataset will be constructed, such that it contains numerical features extracted from the original datasets, indicative of individual's learning engagement in the corresponding platforms. No identifiable data is in the resulting datasets containing the features listed as follows for MOOC data and PeerWise data respectively.

**For MOOC data:**

num_visited_steps	early_begin	pre_AR
num_completed_steps	early_during	pre_FR
precede	n_early_equal	pre_IR
overlap	n_early_finish	pre_LP
during	early_last	pre_SP
abandoned	early_meet	early_AR
equal	early_overlap	early_FR
finish	early_precede	early_IR
meet	late_abandoned	early_LP
last	late_begin	early_SP
num_comments	late_during	late_AR
SP	n_late_equal	late_FR
LP	n_late_finish	late_IR
FR	late_last	late_LP
IR	late_meet	late_SP
AR	late_overlap	post_AR
pre_abandoned	late_precede	post_FR
pre_begin	post_abandoned	post_IR
pre_during	post_begin	post_LP
n_pre_start_equal	post_during	post_SP
n_pre_start_finish	n_post_end_equal	B1
pre_last	n_post_end_finish	B4
pre_meet	post_last	B5
pre_overlap	post_meet	steps_visited_ratio
pre_precede	post_overlap	steps_completed_ratio
early_abandoned	post_precede	eligible_for_certificate

The semantics of these features are detailed as follows:

Feature	Description
<b>SP</b>	Count of starting posts (comments created by the learner which attract replies but are not replies themselves). These are zero-order replies.
<b>LP</b>	Count of lone posts (posts created by the learner which do not attract replies from others and are replies themselves).
<b>FR</b>	Count of first replies (replies to someone else's starting post). These are first-order replies.
<b>IR</b>	Count of initiator replies (replies to someone's reply to their own starting post). These are second-order replies.
<b>AR</b>	Count of additional replies (replies to a reply to a starting post created by someone else). These are also second-order replies.



Feature	Description
<b>precede</b>	Count of steps of Event_type = 'precede'
<b>overlap</b>	Count of steps of Event_type = 'overlap'
<b>during</b>	Count of steps of Event_type = 'during'
<b>abandoned</b>	Count of steps of Event_type = 'abandoned'
<b>equal</b>	Count of steps of Event_type = 'equal'
<b>begin</b>	Count of steps of Event_type = 'begin'
<b>finish</b>	Count of steps of Event_type = 'finish'
<b>meet</b>	Count of steps of Event_type = 'meet'
<b>last</b>	Count of steps of Event_type = 'last'

Feature	Description
<b>B1</b>	One, if the learner has posted at least a comment, zero if not.
<b>B4</b>	One, if the learner has posted at least a first reply, zero if not.
<b>B5</b>	One, if the learner has posted at least an initiators' reply, zero if not.

Feature	Description
<b>steps_visited_ratio</b>	Count of steps visited by the learner over the total number of steps in the MOOC (listed in Table 5.2). Engineered from <code>step-activity.csv</code>
<b>steps_completed_ratio</b>	Count of steps completed by the learner over the total number of steps in the MOOC. Engineered from <code>step-activity.csv</code>
<b>eligible_for_certificate</b>	Calculated as 'True' if their <code>steps_completed_ratio</code> is greater than 0.5. 'False' otherwise.
<b>archetype</b>	Self-reported learning archetype from those listed in Table 2.4 (engineered from <code>archetype-survey-responses.csv</code> when available).

Also, an engineered a final set of features related to bins according to when in the course they were performed. Hence, there are **pre-** and **post-** features, capturing counts of each type of learner activities before and after the formal start and end of the course; there are **early-** features, capturing learner activities in the first ten days of the course; and finally, there are **late-** features, capturing learner activities after the tenth day but before the course ended.

Similarly, for PeerWise data:

Feature	Type	Alternative name	Description
<i>User_ID</i>	string		unique identifier in PeerWise for a student <i>s</i>
MILESTONE BADGES (CAN BE EARNED ONLY ONCE)			
<i>B1</i>	numeric	Question author	<i>s</i> contributed one question
<i>B2</i>	numeric	Question answerer	<i>s</i> answered one question
<i>B3</i>	numeric	Star-crossed	<i>s</i> agreed or disagreed with a comment
<i>B4</i>	numeric	Comment	<i>s</i> wrote one comment
<i>B5</i>	numeric	Author-reply	<i>s</i> replied to a comment written about own question
<i>B6</i>	numeric	Follower	<i>s</i> followed one or more authors
<i>B18</i>	numeric	Leader	<i>s</i> had one or more followers
<i>B19</i>	numeric	Helper	<i>s</i> responded to one help request or more
<i>B23</i>	numeric	Verifier	<i>s</i> has confirmed one answer or more
BADGES THAT CAN BE EARNED MORE THAN ONCE			
<i>B7</i>	numeric	Good question author	per question authored rated as excellent five times or more
<i>B8</i>	numeric	Popular question author	per question authored that was answered ten times or more
<i>B9</i>	numeric	Discussed question author	per question authored that received two or more comments
<i>B10</i>	numeric	Commentator	<i>s</i> wrote five comments or more
<i>B11</i>	numeric	Critic	<i>s</i> agreed or disagreed with ten comments
<i>B12</i>	numeric	Rater	<i>s</i> submitted a rating for ten questions
<i>B13</i>	numeric	Scholar	<i>s</i> answered ten questions correctly
<i>B14</i>	numeric	Genius	<i>s</i> answered ten questions in a row correctly
<i>B15</i>	numeric	Einstein	<i>s</i> answered twenty questions in a row correctly
<i>B16</i>	numeric	Insight	<i>s</i> wrote two or more comments that are agreed with by someone
<i>B17</i>	numeric	Conversation	<i>s</i> replied to five comments about own questions
<i>B24</i>	numeric	Super scholar	<i>s</i> answered correctly a total of 50 questions
TIME SENSITIVE BADGES			
<i>B20</i>	numeric	I'll be back	<i>s</i> answered correctly ten or more questions, on each of three different days)
<i>B21</i>	numeric	Commitment	<i>s</i> answered correctly ten or more questions, on each of five consecutive days
<i>B22</i>	numeric	Obsessed	<i>s</i> answered correctly ten or more questions, on each of ten consecutive days
<i>B25</i>	numeric	Legend	<i>s</i> submitted a correct answer on 31 distinct days

Feature	Data type	Description	Example/ Values
<i>Questions_made</i>	numeric	number of questions (MCQs) authored by <i>s</i>	0...20
<i>Comments_received</i>	numeric	number of comments received by <i>s</i>	0...21
<i>Starting_questions</i>	numeric	number of MCQs authored by <i>s</i> that receive comments	0...12
<i>Lone_questions</i>	numeric	number of MCQs authored by <i>s</i> that do not receive comments	0...8
<i>Comments_made</i>	numeric	number of comments made by <i>s</i>	0...25
<i>Replies_made</i>	numeric	number of replies made by <i>s</i>	0...14
<i>Initiators_Replies</i>	numeric	number of replies made by <i>s</i> to comments on <i>s</i> 's MCQs	0...12
<i>Followers</i>	numeric	number of students who follow <i>s</i>	0...3
<i>Following</i>	numeric	number of students followed by <i>s</i>	0...3
<i>Ratings_given</i>	numeric	number of times that questions <i>s</i> have been rated for quality	0...82
<i>Avg_qual_ratings_given</i>	numeric	average quality rating given to questions by <i>s</i>	0...5
<i>Answers_given</i>	numeric	number of MCQs answered by <i>s</i> (all attempts)	0...529
<i>0-Early_engagement_Question</i>	numeric	As per <i>Questions_made</i> but disaggregated by period	
<i>1-Easter_Question</i>	numeric		
<i>2-Exam_revision_Question</i>	numeric		
<i>3-Post_exam_Question</i>	numeric		
<i>0-Early_engagement_Answer</i>	numeric	As per <i>Answers_given</i> but disaggregated by period	
<i>1-Easter_Answer</i>	numeric		
<i>2-Exam_revision_Answer</i>	numeric		
<i>3-Post_exam_Answer</i>	numeric		
<i>0-Early_engagement_Comment</i>	numeric	As per <i>Comments_made</i> but disaggregated by period	
<i>1-Easter_Comment</i>	numeric		
<i>2-Exam_revision_Comment</i>	numeric		
<i>3-Post_exam_Comment</i>	numeric		
<i>0-Early_engagement_Ratings</i>	numeric	As per <i>Ratings_given</i> but disaggregated by period	
<i>1-Easter_Ratings</i>	numeric		
<i>2-Exam_revision_Ratings</i>	numeric		
<i>3-Post_exam_Ratings</i>	numeric		
<i>0-Early_engagement_Reply</i>	numeric	As per <i>Replies_made</i> but disaggregated by period	
<i>1-Easter_Reply</i>	numeric		
<i>2-Exam_revision_Reply</i>	numeric		
<i>3-Post_exam_Reply</i>	numeric		
<i>Exam_Mark_nominal</i>	enum	Classification from exam marks	first...
<i>Final_Mark_nominal</i>	enum	Classification from final marks	first...



**13. How will you minimise the risk that data subjects (individuals or organisations) could be identified in your presentation of results?**

*Please consider whether disclosive ID codes have been used (e.g. date of birth) and whether it is theoretically possible to identify individuals by combining characteristics (e.g. widow in Hampshire with 14 children) or by combining datasets. How will you protect individuals' anonymity in your analysis and dissemination?*

As mentioned in the amended part of question 10, it may be possible to identify individuals through the combination of the datasets (e.g. comments files, survey responses). However, I will minimise the risk by pre-processing the free-text in the comments files and extracting numerical features indicative of the learning engagement. After this, I will delete the free text from the dataset.

With regards to the survey responses, I will only filter content related to course engagement (and disregard any other that may make reference to personal characteristics). In reporting my findings, I would report group characteristics (rather than individuals), and, if appropriate, quote any survey responses after stripping the anonymised userID of the learner assigned by FutureLearn, and assigning it a new one (e.g. "Learner A").

The above mentioned risk does not exist with the part of the study involving students in face-to-face instruction, as the only personal information used is the student ID, which is used for the purpose of linking the two datasets (for PeerWise and QMP) but it is replaced with another unique identifier when pre-processing the data. This effectively means that the link to the original identity of each individual will be destroyed before the data analysis stage.

**14. What other ethical risks are raised by your research, and how do you intend to manage these?**

*Issues may arise due to the nature of the research you intend to undertake and/or the subject matter of the data. Examples include: data or analysis that are culturally or socially sensitive; data relating to criminal activity, including terrorism, and security sensitive issues.*

none

**15. Please outline any other information that you feel may be relevant to this submission.**

*For example, will you be using the services or facilities of ONS, ADRN, or HSCIC and/or are you obtaining ethical review from NRES (through IRAS) or other? Please confirm whether the data being used are already in the public domain.*

None other than those explained above.

**16. Please indicate if you, your supervisor or a member of the study team/research group (including any institution that they act for, if different from the University) are a data controller and/or data processor in relation to the personal data you intend to process as defined by the Data Protection Act 2018 following the GDPR, and confirm that you/they understand your/their respective responsibilities ( <https://ico.org.uk/for-organisations/guide-to-data->**

[protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/controllers-and-processors/\)](#)

I am both a data controller and data processor, as follows:

- For the *PeerWise* data: I have been a jointly data controller with Paul Denny, from the University of Auckland in New Zealand. Paul is the creator of PeerWise and as such, he holds the dataset associated with the data-subjects participation in the COMP2213 course within *PeerWise*.
- For the *QuestionMark Perception* data: I have been a jointly data controller with Steve Snow (ss33g15, for both cohorts under consideration), m.c. schraefel (mc, for the 2015/16 cohort only), and Nick Gibbins (nmg, for the 2016/17 cohort only), as we were all examiners and had access to the exam files.

I am the only data processor with access to both sets of data and I fully understand my responsibility to de-anonymise the combined data prior the creation of the feature vector (and qualitative analysis of the survey data) as explained in points 7, 10 and 13.

## Guidance on applying for ethics approval for secondary data analysis

If your research PURELY involves the following, you do not need to apply for ethics approval:

- analysis of aggregated data on individuals or organisations (e.g. GDP, labour force participation rates, fertility rates);
- meta-analyses (i.e. the analysis of studies);
- literature reviews or reviews/analyses of reports, policies, documents, meeting minutes, newspaper articles, films.

### Filling in the online submission form on ERGO II:

- Please give your application a title that includes 'SDA' (Secondary Data Analysis).
- Please refer to the "**ERGO II Guidance for Applicants**" document (downloadable from the ERGO II site) on how to answer the submission questionnaire correctly..

### Additional Forms:

If your study PURELY involves secondary analysis of data, you only need to fill in the 'Ethics Application Form for Secondary Data Analysis'. You do not need a Risk Assessment Form.

If your study is a mixed-method study involving secondary data analysis AND some component of data collection (e.g. interviews, online survey), then you need to fill in additional forms:

- Ethics Application Form (for studies other than secondary data analysis)
- Risk Assessment Form
- Participant Information Sheet
- Consent Form
- Draft research instrument

### Please note:

- You must not begin data analysis until ethical approval has been obtained.
- It is your responsibility to follow the University of Southampton's Ethics Policy and any relevant academic or professional guidelines in the conduct of your research. This includes ensuring confidentiality in the storage and use of data.
- It is your responsibility to provide full and accurate information in completing this form.

# Appendix **B**

## **Data Protection Impact Assessment**

The following are the documents related to the Data Protection Impact Assessment (DPIA), associated to iSolutions ticket [RITM0296306]. This was approved by the DPIA panel at the University of Southampton on session dated 9th July 2020, as per Part C of the report.

### **DATA PROTECTION IMPACT ASSESSMENT**

Data protection impact assessments (DPIAs) are a risk assessment that help the University identify the most effective way to comply with its data protection obligations and meet individuals' expectations of privacy whilst also allowing it to identify and fix problems at an early stage.

You have completed an Initial Data Protection Review, which has highlighted that the project/programme you are proposing involves the processing of data relating to people from which they can be singled out, or where the processing to be carried out under the project could potentially harm individuals and/or negatively impact their rights and freedoms ('personal data processing').

The University takes a risk based approach to data protection compliance. So as to enable the University to assess the level of risk associated with your project/programme **please complete Part A** below and attach any relevant documents in support, including the Initial Data Protection Review and the project proposal, into Servicenow.

Once you have completed **Part A**, the Data Protection Impact Assessment Panel will review it and complete **Part B** (Privacy Impact Action Plan) and **Part C** (Authorisation). The Panel will then send it back to the Board and yourself with any recommendations and for your assistance in completing any of the actions identified in Part B. Please note that the DPIA will need to be reviewed on a regular basis.

Title of activity	Learner Engagement within Peer-Supported Environments mediated by Digital Technologies: PeerWise and MOOC Platforms
Brief Background/ description of activity	The rationale of this PhD project lies on the need to understand the patterns of learner engagement using digital technologies in peer-supported environments, such as with PeerWise in the context of face-to-face instruction, and also in the context of massive open online courses (MOOCs). The data related to this DPIA form is restricted to two MOOCs by the University of Southampton .
Explain broadly what your activity aims to achieve.	<p>This research uses secondary data analysis, on data collected by FutureLearn on the University of Southampton MOOCs "Understanding Language" and "Portus" (ERGO/FEPS/55694). This dataset contains activity data of all of the learners who undertook these courses.</p> <p>I intend to process the engagement data available across various files, and generate a learner "profile" which includes, for each week of the course, the number of articles read, number of videos viewed, discussions joined, number of completed assignments, generated comments and likes received. Other files of this dataset include survey responses within the platform which are also valuable to judge the learners' engagement and attainment.</p> <p>In generating these profiles, I am only wishing to observe clusters of behaviour across the learner population, therefore, such "profiling" will not (and cannot) be used to impact any individuals. However, I acknowledge that as part of the analysis I might encounter some personal data which may have disclosed as part of the comments written by themselves within the learning activities of these two courses under study.</p> <p>Other data within the dataset which is considered personal is that within the "enrolments" files, in which I am not interested and do not wish to receive or process.</p>

B-3

Key Project Contacts:	Name	Job title	Email
Researcher	Adriana Wilde	PhD researcher	agw106@soton.ac.uk
Supervisor	David Millard	Associate Professor	dem@soton.ac.uk

**PART A: ASSESSMENT**

**1. Data must be processed lawfully, fairly and in a transparent manner.**

<p><b>A.</b></p>	<p>Brief description of the type of personal data being processed and what it will be used for:</p>	<p>I, the researcher, as part of my PhD thesis “<b>Clustering Analysis of Learner Engagement within Peer-Supported Environments mediated by Digital Technologies: PeerWise and MOOCs Platforms</b>”, aim to identify the main patterns of engagement using a specific machine learning technique that involves the grouping of individual data points. In addition to the description of engagement clusters, the data will be used to predict attainment. In general terms, the research aims to identify whether a specific pattern of engagement is more indicative than others to achieve academic success, i.e. higher marks in the case of face-to-face instruction (another part of my study, done over a different dataset) or, in the case of MOOCs, retention and completion. My method for processing the data is in preprocessing the original dataset so that <b>any profiling of individuals within a group will not lead to their identification</b>. The dataset will be anonymised and <b>the profiling will be applied only on the anonymised data</b>. I will remove any identifiable data (including the original identifiers, and free text comments) and therefore, there will be no risk of identifying any individuals. In my dissemination of findings, <b>I will report on clusters of behaviour of anonymous individuals</b>.</p>
<p><b>B.</b></p>	<p>What entitles you to process the personal data i.e. what is your lawful basis for processing?</p>	<p><input type="checkbox"/> Clear consent has been/will be obtained</p> <p><input type="checkbox"/> It is necessary for compliance with a legal obligation to which the University is subject</p> <p><input type="checkbox"/> It is necessary for the performance of a contract to which the data subject is a party</p> <p><input type="checkbox"/> It is necessary to protect the individual’s vital interest (to protect someone’s life)</p> <p><input checked="" type="checkbox"/> It is necessary for the performance of a task carried out in the public interest or for your official functions and the task/function has a clear basis in law</p> <p><input type="checkbox"/> The processing is necessary for the University’s legitimate interests or the legitimate interests of a third party*</p>
	<p>Please provide details as to why you have chosen this ground as your lawful basis. *If you are relying on the legitimate interest basis please complete the Legitimate Interest Assessment at Appendix A and attach it to the DPIA.</p>	<p>As a university, we are expected to train students and researchers, and to carry out quality research. That is our remit.</p>

<p><b>C.</b></p>	<p>Are you processing any <b>Special Category (sensitive) Personal Data</b>? If so, what is your lawful basis for processing?</p> <p>NB: to process Special Category Data you must be able to satisfy one of the grounds set out in <b>B</b>, as well as one of the grounds set out in <b>C</b>.</p>	<p><input type="checkbox"/> Explicit consent has been <del>will</del> be obtained</p> <p><input type="checkbox"/> Necessary for carrying out obligations under employment or collective contract</p> <p><input type="checkbox"/> Necessary to protect the data subject's vital interests</p> <p><input type="checkbox"/> Necessary for medical treatment, assessing working capacity, provision off social care or a contract with a health professional</p> <p><input type="checkbox"/> Necessary for reasons of public interest in the area of public health</p> <p><input type="checkbox"/> The data has been manifestly made public by the data subject</p> <p><input type="checkbox"/> Necessary for the establishment, exercise or defence of a legal claim</p> <p><input type="checkbox"/> Necessary for reasons of substantial public interest under UK law</p> <p><input checked="" type="checkbox"/> Necessary for archiving purposes in the public interest or <b>scientific</b> and historical <b>research</b> or statistical purposes</p>
	<p>Please provide details as to why you have chosen this ground as your lawful basis: (Please see attached guidance notes)</p>	<p>This is necessary for scientific research purposes as detailed in Question A.1.A. above.</p>
<p><b>D.</b></p>	<p>If consent has been obtained, where has this been recorded?</p>	<p>Upon registration on the concerned FutureLearn courses, participants consent to their use of data to perform academic research (see para 4.1.k and para 12.3 of FutureLearn's Terms and Conditions and <a href="#">Privacy Policy</a>). Please note in particular:</p> <p><b>12.3</b> Your activities on an Online Course are shared with the course provider for academic research purposes. This includes the comments you make where you may disclose certain personal information about yourself.</p>
	<p>If relying on consent, what will happen if consent is withheld/ withdrawn?</p>	<p>If consent is withheld/withdrawn before the end of my project, the associated data will be deleted from the dataset. This would not be possible once the findings have been published in the final thesis or subsequent academic publications, however, no data leading to the identification of the learner would ever be published.</p>
	<p>If consent is obtained, what processes are in place to enable the data to be erased if requested?</p>	<p>I would require FutureLearn to let me know the <i>LearnerID</i> of the participant (described in A.1.G., below. With this information I can easily delete all entries related to that <i>LearnerID</i> across my processed dataset (if the request arises prior the end of my research, as explained above).</p>



	How will you ensure any obligations are complied with?	GDPR obligations are embedded throughout the design process as reported in the approved ethics application <b>ERGO/FEPS/55694</b> .
<b>E.</b>	Is there any data being processed/held under a duty of confidentiality?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No Please provide details:
<b>F.</b>	Will a 3 <sup>rd</sup> party process the data on the University's behalf?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No Please provide details:
<b>G.</b>	Will data be anonymised or pseudonymised for processing?	<input checked="" type="checkbox"/> Yes* Please provide details: The data I am requesting is already in anonymised form. Rather than names or usernames, the unique identifier per participant is a FutureLearn-issued <i>LearnerID</i> , of the form xxxxxxxx-xxxx-xxxx-xxxx-xxxx, where x is either a number or a letter. However, there is a risk that a participant may have disclosed sensitive information as part of their comments or in survey responses, as mentioned in A.1.D., above. In this case, I will mitigate this risk by deleting any information of such kind that I might encounter during preprocessing, as pledged to in the ERGO/FEPS/55694 document and per FutureLearn assurances to learners in para 12.4 of their <a href="#">Privacy Policy</a> : “we confirm that all our course providers who conduct research will never associate your comment or your activity with your user account by method (b) above and will always treat any personal data in strict accordance with data protection laws and the research ethic guidelines.”
<b>H.</b>	Is the data in respect of vulnerable individuals? No	Category: N/A Special arrangements required: N/A

**2. Data collected for specified, explicit and legitimate purposes and not processed further for any incompatible purpose(s).**

<b>A.</b>	How are you collecting the data?	It is a secondary data analysis research. It has already been collected by FutureLearn and shared with the University of Southampton as a partner and course provider for these MOOCs.
<b>B.</b>	Who will retain responsibility for the personal data under the Data Protection Act? (i.e. Who is the Data Controller?)	As it is a secondary data analysis research, the data controller is the one defined at the point of data collection, i.e. the University of Southampton as course provider and FutureLearn, as per FutureLearn's <a href="#">Privacy Policy</a> , para. 1.1.: FutureLearn Limited, a company incorporated in England and Wales (registered number 8324083) whose registered office is at 1-11 Hawley Crescent, Camden Town, London, NW1 8NP, United Kingdom, [...] will be the controller of any personal data processed as described in this <a href="#">Privacy Policy</a> .
<b>C.</b>	Will it be collected directly from the data subject or from a third party? If it is a third party, please give details.	See A.2.A., above for collection from the data subject, and A.4.A., below, for details on the third party: Ms Chrissie Metcalf, iSolutions (University of Southampton).
<b>D.</b>	How will individuals (data subjects) be made aware of whom the Data Controller is?	They became aware on registration to these courses, and confirmed having read and accepted the terms and conditions (including the <a href="#">privacy policy</a> , and the paragraph shown in answer A.2.B. above) by clicking on a check box. Without having done so, they would not have been able to take these courses, and therefore, none of their data would have been collected.
<b>E.</b>	Will you be using any other organisation to process data on your behalf?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No Please provide details:
<b>F.</b>	Will the individuals (data subjects) still have control over their data?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No Please provide details:

<b>G.</b>	What are the reasons for collecting the data?	Academic research as explained in question A.1.A., above
<b>H.</b>	How will you use the data (if different from above)?	N/A
<b>I.</b>	Are the individuals/data subjects likely to expect this processing to be taking place?	Yes, because this is one of the purposes of the collection/storage/use/sharing of their data, as outlined in FutureLearn's <a href="#">Privacy Policy</a> , para. 4.1.(k):  to allow universities and other partner institutions that provide Online Courses or Content to perform academic research (as more fully set out in paragraph 12 below);  and in para. 12.3:  Your activities on an Online Course are shared with the course provider for academic research purposes. This includes the comments you make where you may disclose certain personal information about yourself.
<b>J.</b>	How will collecting this data benefit staff, students and/or the University?	This research is important to inform the university how best provide peer-supporting environments mediated with technology (such as MOOCs), and any findings of my research will be part of continuing quality-improvement processes at the university. Additionally, as per FutureLearn Terms and Conditions and <a href="#">Privacy Policy</a> , para 12.1.:  The universities and other partner institutions that provide Online Courses or Content on the Website carry out academic research as they are dependent on research funding in order to develop and provide Online Courses.
<b>K.</b>	Will this Project interfere with the data subjects' right to privacy under the Human Rights Act?	No
<b>L.</b>	How and when will data subjects be told where and why their data is being held, and how it will be used? Are you issuing a Privacy Notice?	This type of research is already covered by FutureLearn's <a href="#">Privacy Policy</a> as explained above. Therefore, not only it is unnecessary for me to issue it, I

		would not be able to communicate it to the data subjects as I will not have their email addresses.
<b>M.</b>	How and when will data subjects be given the option to 'opt out' or restrict processing for other purposes?	On registration to FutureLearn courses, data subjects are told they can opt out at anytime by emailing <a href="mailto:support@futurelearn.com">support@futurelearn.com</a> (and they can also "access, correct, or update" their information), as explained in section 9.1. of FutureLearn's <a href="#">Privacy Policy</a> . In such eventuality, I will then need to receive from FutureLearn their <i>LearnerID</i> (described in A.1.G., above)
<b>N.</b>	Will any decisions affecting data subjects be made solely on processing by automatic means?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No Please provide details:
<b>O.</b>	If new purposes are identified, how will data subjects be informed of this?	N/A

B-9

**3. Data is adequate, relevant and limited to what is necessary for the purposes for which they are being processed.**

<b>A.</b>	What is the media being used (e.g. paper records, electronic etc.)?	Electronic records in the form of comma-separated-value files.
<b>B.</b>	What is the volume of the data being processed (e.g. all staff records, small research cohort of 20 students)?	Tens of thousands of learners. The complete dataset consists of the following files, presented in tables 1.16 and 1.17, presented in page 9, below. Each table refers to the number of data entries associated to each course. Note that from this set I will require only the files listed under A.6.K, below

TABLE 1.16: Summary table of entries per file in the Portus MOOC dataset per run

file	Run					
	1	2	3	4	5	6
enrolments	7773	8920	3252	5172	4266	1286
step-activity	281159	213537	58559	94904	84356	41964
comments	20253	18846	3566	13929	12465	5010
question-response	133749	100842	26840	47329	43477	21821
peer-review-assignments	265	356	89			
peer-review-reviews	659	681	109			
weekly-sentiment-survey-responses						9
video-stats						73
archetype-survey-responses						157
leaving-survey-responses						85

TABLE 1.17: Summary table of entries per file in the Understanding Language MOOC dataset per run

file	Run										
	1	2	3	4	5	6	7	8	9	10	11
enrolments	58782	41913	44284	25591	19873	10279	12900	6034	8311	5096	7832
step-activity	467333	317324	228623	197266	127584	73588	115204	41149	74832	41564	61110
comments	145425	86139	58285	50332	37637	18616	24941	9307	16368	8469	13229
archetype-survey-responses							1586	607	946	503	798
leaving-survey-responses							185	128	137	78	128
weekly-sentiment-survey-responses							140	232	116	205	
video-stats							32	35	35	35	
post-course-survey-data											166
post-course-survey-free-text											64

<b>C.</b>	How will the data be assessed for relevance to ensure that no more than the minimum required is collected?	Scoping it the to specific courses and files requested (see A.6.K, below). These are the only relevant files for studying the engagement and attainment of learners. No more than these are being requested.
<b>D.</b>	What is the envisaged extent and frequency of the processing?	One-off process, for the analysis section within my PhD thesis
<b>E.</b>	How, when and by whom will checks be conducted to ensure only the minimum is being held?	Myself with my supervisor, as discussed and approved by the Faculty Ethics Committee (ERGO/FEPS/55694)
<b>F.</b>	Will the quality and quantity of the data be sufficient for the intended purpose?	<input checked="" type="checkbox"/> Yes* Please state why: The model of the learner engagement will be constructed to compare against that of student engagement in face-to-face instruction using PeerWise. In both cases, digital traces of engagement are captured across the whole cohort of each course.
<b>G.</b>	Is there any data you could avoid using? Please provide details.	<input checked="" type="checkbox"/> Yes* The full dataset that FutureLearn has given to the University of Southampton includes enrolment data (which includes demographic data, such as gender, country, age range, employment status, etc). I DO NOT need nor want to receive these files as I can avoid using them. Similarly, though I do need to process data from the comments files and survey responses files to measure engagement and attainment, I will not use any text which may contain personal data. Such text would be removed as part of the pre-processing.  <input type="checkbox"/> No

4. Accurate and, where necessary, kept up to date.

<p><b>A.</b> Will the data be received from third parties?</p>	<p><input type="checkbox"/> Yes*</p>	<p><input type="checkbox"/> No</p>
<p>If yes, how will it be received?</p>	<p><input type="checkbox"/> On paper</p>	<p><input type="checkbox"/> Other</p>
<p><b>B.</b> Will the data be received directly from the data subject?</p>	<p><input type="checkbox"/> Verbally</p>	<p><input checked="" type="checkbox"/> Electronically</p>
<p>If yes, how will it be received?</p>	<p><input type="checkbox"/> Yes*</p>	<p><input checked="" type="checkbox"/> No</p>
<p><b>C.</b> Will the data be sent to third parties?</p>	<p><input type="checkbox"/> On paper</p>	<p><input type="checkbox"/> Other</p>
<p>If yes, how will it be provided?</p>	<p><input type="checkbox"/> Verbally</p>	<p><input type="checkbox"/> Electronically</p>
<p><b>D.</b> Will the data be sent out to the data subjects?</p>	<p><input type="checkbox"/> Yes*</p>	<p><input checked="" type="checkbox"/> No</p>
<p>If yes, how will it be provided?</p>	<p>Please provide details:</p>	<p><input type="checkbox"/> On paper</p>
<p><b>E.</b> Will the data link with other existing or new systems/processes?</p>	<p><input type="checkbox"/> On paper</p>	<p><input type="checkbox"/> Other</p>
<p>If yes, how will it be provided?</p>	<p><input type="checkbox"/> Verbally</p>	<p><input type="checkbox"/> Electronically</p>
<p><b>F.</b> How will the data be stored in the University?</p>	<p><input type="checkbox"/> Yes*</p>	<p><input checked="" type="checkbox"/> No</p>
<p><b>G.</b> Where will it be stored? Please include details of backups and copies.</p>	<p>Please provide details:</p>	<p><input type="checkbox"/> Paper &amp; Electronic</p>
<p><b>G.</b> Where will it be stored? Please include details of backups and copies.</p>	<p><input type="checkbox"/> Verbally</p>	<p><input checked="" type="checkbox"/> Electronically</p>
<p>University storage, accessed via a university-provided laptop, regularly backed up by the university. I receive all my secondary data via <b>safesend.soton</b>, use <b>git.soton.ac.uk</b> for version control of the code performing the preprocessing of the data, and use <b>OneDrive for Business/Sharepoint</b> for secure storage of the secondary data as received as well as all the LaTeX files associated with my thesis (including all the bibliography used, in the form of .TeX and .bib files). This means that iSolution will be able to assist with recovering any accidentally</p>		

		deleted files (for up to 90 days after deletion), that the files are encrypted at rest, and that all data is held in secure centers within the UK.
<b>H.</b>	What procedures will be in place to ensure it is kept up to date/accurate?	As it is all participation data on finished courses, I will not require to update it or modify it other than for protecting the identity of the participants.
<b>I.</b>	How and when will the data be checked for accuracy and completeness?	This data will be received by me directly from Chrissie Metcalf from iSolutions, who has created the processes to manage GDPR compliance of MOOC data for research, so will not give me inaccurate or incomplete data.
<b>J.</b>	How and when will the accuracy be checked with the data subject?	N/A
<b>K.</b>	If the data subject believes their data is incorrect or incomplete how will this be processed?	As explained in A.1.D., above, I would require FutureLearn to provide me with the <i>LearnerID</i> of the data subject, as well as the actual part of the dataset which would need updating.
<b>L.</b>	What would the risk/impact of using inaccurate/out of date data?	No risk to any individual data subjects. However, if the data is inaccurate, then any insights on the data analysis would, of course, be inaccurate too. Out of date data is somewhat expected as the analyses are done over completed courses rather than ongoing ones. Updating data is not needed in this study.

B-13

**5. Kept for no longer than is necessary.**

<b>A.</b>	What will be the record start- up and close-down procedures?	<p><b>START-UP:</b> Once I receive the dataset via SafeSend from Chrissie Metcalf.</p> <p><b>CLOSE-DOWN:</b> Once the research is finished, the dataset will be deleted from my allocated university storage, but the responsibility will fall to my supervisor to arrange that the data is stored appropriately (as described in my Data</p>
-----------	--	---



		Management Plan).
<b>B.</b>	How long is it envisaged that the data will be retained for? What criteria will be used?	My PhD thesis is to be resubmitted no later than 28 February 2021. I will only require access to the data beyond this point if the examiners are not able to examine the thesis before then or are not yet satisfied with the findings as presented then.
<b>C.</b>	What is the intended process for ensuring that data is not kept for longer than required?	As the storage is associated with my university student account, once I cease to be a student, I will no longer have access to such data.
<b>D.</b>	How will data be managed/disposed of when it is no longer required?	It will be deleted from university storage.
<b>E.</b>	Will the data be stored for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No
<b>F.</b>	What technical or organisational measures are in place to safeguard the rights and freedoms of the data subject? Please provide details.	As explained above, and in compliance to GDPR and honoring the assurances made to data subjects in <a href="#">FutureLearn privacy policy</a> .
<b>G.</b>	Are there any anticipated exceptional circumstances for retaining certain data for longer than normal?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No

6. Information and system security include protection against unauthorised/unlawful processing and accidental loss, destruction or damage, using appropriate technical or organisational measures.

A.	How will the data be accessed?	<input type="checkbox"/> Paper & Electronic	<input checked="" type="checkbox"/> Electronic <input type="checkbox"/> Other
B.	Who will have access to the data?	For the purposes of this DPIA request, just me (Adriana Wilde, the researcher) and my supervisor (Dave Millard), who will only have access to pre-processed data by me (already stripped of identifiable data) in order to offer advice on how to present/discuss my findings. However, as explained earlier, this research is a secondary data analysis, and as such, other people within our organisation may also be granted access rights to the same dataset, so they will need to do their own applications to ERGO, IDPR, DPIA, etc.	
C.	Will access be restricted on the basis on staff roles/seniority?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know	<b>Please provide details:</b> No people other than myself will look at the copies of the requested files. The restriction is not based on staff roles/seniority.
D.	Who will be responsible for providing access rights?	Chrissie Metcalf (Technical Product Manager in Digital Learning, iSolutions).	
E.	Is there an audit trail showing new/amended access rights are granted and by whom?	<input checked="" type="checkbox"/> Yes* <input type="checkbox"/> No <input type="checkbox"/> Don't know	Please provide details: Only myself and my supervisor will have access as described above. In the unlikely event of changing supervisors for any unforeseeable reasons, the Grad school would appoint a new supervisor and I would be notified through auditable university systems Similarly, all the requests for access have been made either via email (to Kate Borthwick, Chrissie Metcalf, etc) or via web forms in iSolutions, using an auditable ticket system (e.g. this form is part of the ticket RITM0296306). Future changes to access rights related to this research would be auditable in the same ways.
F.	What security measures will be in place for the system/process?	By using password-protected, university-stored encrypted files. <input type="checkbox"/> Don't know	
G.	What procedures will be in place for detecting and dealing with security breaches?	Any breaches will be reported to iSolutions, who are entrusted and responsible for protecting the information security throughout the organisation. <input type="checkbox"/> Don't know	
H.	How, when and by whom will the security measures be audited for compliance?	This is the responsibility of iSolutions, who maintain the university systems. <input type="checkbox"/> Don't know	

<b>I.</b>	Will the system include reports?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No <input checked="" type="checkbox"/> Don't know
<b>J.</b>	Who will have access to these reports?	Please provide details including purpose of records: <input type="checkbox"/> University staff only <input type="checkbox"/> Third parties <input type="checkbox"/> Other <input type="checkbox"/> Data subjects <input type="checkbox"/> Don't know <b>N/A</b>
<b>K.</b>	What data will be shared, and with whom?	<p>The data I request (to be shared with me alone) is related to the datasets of all of the 11 runs of the Understanding Language MOOC, and all of the 6 runs of the Portus MOOC. In particular, and as per email to Kate Borthwick and Chrissie Metcalf, dated 27 May 2020:</p> <p>“the unstripped data (i.e. all columns for each file) ONLY for the following:</p> <ul style="list-style-type: none"> <li>• step-activity</li> <li>• comments</li> <li>• archetype-survey-responses</li> <li>• post-course-survey-data</li> <li>• post-course-survey-free-text</li> <li>• weekly-sentiment-survey-responses</li> </ul> <p>note that I am excluding, amongst others, the enrolment files.”</p> <p>In addition: the question-response files from the Portus MOOC (all runs), which do not have any sensitive information. As evidence, here are the first two lines of the files I had received earlier in the year (in which currently the pseudonymised userID is missing, and is therefore unusable for my purposes):</p> <pre>quiz_question,question_type,week_number,step_number,question_number,response,cloze_response,submitted_at,correct 1.20.1,MultipleChoice,1,20,1,5,,2015-06-15 02:15:35 UTC,TRUE</pre>
<b>L.</b>	By what method(s) will the data be shared?	<input type="checkbox"/> Telephone <input type="checkbox"/> Post – internal <input type="checkbox"/> Post external <input type="checkbox"/> System link <input type="checkbox"/> By Hand <input type="checkbox"/> University email <input type="checkbox"/> Website <input type="checkbox"/> Fax <input type="checkbox"/> Don't know <b>ELECTRONICALLY WITH SAFESEND</b>
<b>M.</b>	Are there any known or anticipated data related risks/issues?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know Please provide details in the Additional Information box at the end.

<b>N.</b>	What business continuity measures will be in place for the system/process?	N/A	
<b>O.</b>	What is the risk management process for recovering data, which may be damaged/lost?	As I am requesting a copy of these datasets, there is no additional risk: the University and FutureLearn have already processes in place for protecting/recovering the original data in case of it being damaged or lost, as explained in paragraph 5.2 of FutureLearn's <a href="#">privacy policy</a> : "We have put in place technical and organisational security measures to prevent the loss or unauthorised access of your personal information."	
<b>P.</b>	What training/advice/guidance will be available to staff on the new system/process?	N/A	
<b>Q.</b>	Is there other work/other Projects taking place which will (or could potentially) impact on this Project?	<input type="checkbox"/> Yes* Please provide details:	<input checked="" type="checkbox"/> No
<b>R.</b>	Are there any current issues of public concern about this type of processing?	<input type="checkbox"/> Yes* Please provide details:	<input checked="" type="checkbox"/> No
<b>S.</b>	Are there to be any Codes of Conduct or stakeholder requirements applicable e.g. NHS Data and Privacy management requirements?	<input checked="" type="checkbox"/> Yes* Please provide details: Those outlined in the Ethical Research Governance guidelines of the University of Southampton and FutureLearn's <a href="#">privacy policy</a> mentioned throughout this document.	<input type="checkbox"/> No

B-17

**7. Data Subject rights – What procedures are in place to ensure the data subject can:**

<b>A.</b>	Have the right to be informed?	Yes, they do, through FutureLearn. This research is secondary data analysis, therefore it does not infringe on this right. The procedures in place ensure that they are upheld as per the data subjects' agreement with FutureLearn when they enrolled on their courses (see section 9 on the <a href="#">Privacy Policy</a> ).
<b>B.</b>	Request access to their own data?	<p>Yes, they do, through FutureLearn. As I do not collect any additional data from the data subject, they can request all/any of the data I would hold of them, directly to FutureLearn. The data subject must write their request to the following address (as specified in the <a href="#">Privacy Policy</a>):</p> <p>FutureLearn Limited          FAO: Data Protection Officer          1-11 Hawley Crescent          Camden Town          London NW1 8NP          United Kingdom          E-mail: <a href="mailto:support@futurelearn.com">support@futurelearn.com</a></p>
<b>C.</b>	Request data be erased and processing prevented where specific criteria are met?	Yes, as explained in A.7.B, above. Once the data subject makes such request to FutureLearn, I would be able to be notified (via the University of Southampton) with the <i>LearnerID</i> of the data subject so I can erase their data from their dataset. This request would be fulfilled only if made before my thesis is completed.
<b>D.</b>	Request restricted processing of the data where certain circumstances apply?	Yes, as explained in A.7.B, above.
<b>E.</b>	Request portability of the data for their own purpose (if applicable)?	Yes, the data is portable (they are all comma-separated-values files, CSVs). They can do these requests as explained in A.7.B, above.
<b>F.</b>	Object to processing for direct marketing and for the purposes of scientific/historical research and statistics?	I would not use the data for direct marketing purpose anyway, therefore, their objecting to this use would not prompt any additional actions by me. Should a data subject object to the processing of their data for the purpose of scientific research, they will need to notify FutureLearn, as explained in A.7.B, above. Then I would need to be notified via the University of Southampton so that I can delete that data from the dataset (as explained in A.7.C, above).
<b>G.</b>	Exercise right not to be subject to a decision based on an automated process or profiling?	I would not use their data with the purpose of subjecting anyone to any decision anyway. So their objecting would not prompt any additional actions by me.

<b>H.</b>	Rights in relation to international transfers?	I would not use their data with this purpose anyway. So their objecting would not prompt any additional actions by me.
<b>I.</b>	Rights in relation to prior consultation?	Prior consultation was done by FutureLearn as outlined in FutureLearn's <a href="#">Privacy Policy</a> , para. 4.1.(k) and explained in A.2.I, above.
<b>J.</b>	Rights in relation to automated decision-making and profiling?	<p>I would not use their data for automated decision-making anyway. So their rights would always be exercised without prompting any additional actions by me. With regards to "profiling", as explained in the approved ERGO/FEPS/55694 document, the features used for creating the learning profile are to characterise learning engagement rather than any special characteristics or personal data. Excerpt included here:</p> <p>"For this part of the study, the engagement data available across various files allows for a learner profile which includes, for each week of the course, the number of articles read, number of videos viewed, discussions joined, number of completed assignments, generated comments and likes received. Other files of this dataset include survey responses within the platform which are also valuable to judge the learners' engagement and attainment."</p> <p>Having said this, if FutureLearn (via the University of Southampton) informs me that I should remove a data subject from my study, I will be able to do it (provided I'm given their <i>LearnerID</i>).</p>

B-19

### 8. International Transfers

<b>A.</b>	Where is the data being transferred to?	<input checked="" type="checkbox"/> Only internally within University <input type="checkbox"/> Third parties in UK <input type="checkbox"/> Third parties outside EEA *
	*To what countries outside of the EEA will data be transferred to?	N/A
	*What data is being sent to these countries?	N/A
<b>B.</b>	Location of Data Subjects	Unknown by me – potentially anywhere in the world.
<b>C.</b>	Location of Users (employees, contractors. Third parties)	N/A
<b>D.</b>	Hosting location	University of Southampton storage
<b>E.</b>	Support and maintenance (application support and maintenance)	iSolutions

<b>F.</b>	Country specific documents	N/A
<b>G.</b>	International transfer arrangements	N/A <input type="checkbox"/> Don't know
<b>H.</b>	Name and role of persons receiving the data	Only the researcher, Adriana Wilde, <a href="mailto:agw106@soton.ac.uk">agw106@soton.ac.uk</a>
<b>I.</b>	Grounds for transfer (Corporate rules, Model clauses, privacy shield)	<input checked="" type="checkbox"/> Don't know
<b>J.</b>	What are the risks involved in sending this data?	All risks are mitigated by using SafeSend for data transfer, which allow for end-to-end encryption and password protection.
<b>K.</b>	Are the data subject's rights and remedies still enforceable following the transfer?	<input type="checkbox"/> Yes* <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know
<b>L.</b>	Have you sought advice from Legal Services as to whether adequate safeguards exist or whether any other derogations apply?	<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Please provide details:
<b>M.</b>	What ongoing measures are in place to monitor safeguards?	Those in place and managed by iSolutions.

B-20

**9. Disclosure to third parties (internal & external) – Stakeholders & statutory/regulatory bodies e.g. Office for Students, Home Office etc. (add as appropriate)**

Recipients (add their ICO number if applicable)	1.	2.	3.
Name	Dr David Millard		
Address	University of Southampton, <b>Building 32</b> , Level 3, Room 3031L.		
Role	PGR Supervisor		
Data to be disclosed	Pre-processed data (after free-text within the dataset has been stripped of any sensitive data)		
Role of the recipient	PGR Supervisor		
Reasons for disclosure	To give adequate supervision of this research, particularly with guidance on performing analysis and in presenting the findings		

Is there a data sharing agreement/contract in place?	As per the Graduate School regulations and the responsibilities agreed with the University in the employment contract.		
Need for separate for DPIA's	no		
Monitoring arrangements/contract management	Regular supervisory meetings. Report with the Graduate School when required.		

### 10. Additional information

Please use this space to provide any additional information which is relevant to the privacy of the data including any concerns it has highlighted or questions you may have:	I would like to take to opportunity to stress that even though in my research methodology I am applying machine learning algorithms to build learners "profiles", none of these are attributable to any particular data subject. Further, that my preprocessing of the data (in particular my feature extraction processes) will guarantee that no identifiable data is used, even if present in any of the free-text within the dataset under consideration.
---	---

B-21

### 11. Attachments

Please attach all supporting documents to this DPIA including the Initial Data Protection Review and any:

- Information asset register
- Consideration of the balance test were you are relying on legitimate interest as your lawful basis.
- Retention Schedule and/or policy
- Data flow diagrams

<b>Sign Off:</b>	<b>Project Lead</b>	<b>Project Sponsor/ Other</b>
<b>Signature:*</b>	<i>A. Wilde</i>	D. E. Millard
<b>Full name:</b>	Adriana Gabriela Wilde	Dr David E Millard
<b>Job Title:</b>	PhD researcher in Computer Science	Associate Professor in Computer Science
<b>Date of Approval:</b>	<i>09 June 2020</i> <i>Revised: 09 August 2020</i>	10 June 2020



(THIS PAGE OF THE DPIA REPORT WAS LEFT INTENTIONALLY BLANK)

**PART B: RISK MANAGEMENT TABLE & PRIVACY RISK IDENTIFICATION**

The DPIA Panel completes the Privacy Risk Identification form. It will be returned to the Board and the Project Lead along with the DPIA Reporting Form.

- **DPIA Panel Conclusion: Privacy risk identification form not needed to be completed in this case as assessed lower risk post-review.**  
 Authorised: Dr Alison Knight (Head of Research Integrity & Governance, University of Southampton)

**RISK MANAGEMENT TABLE**

Impact	Financial (comparative purposes only)	Legal (comparative purposes only)	Health & Safety (comparative purposes only)	Business Continuity (comparative purposes only)	Reputational (comparative purposes only)
1 Trivial - insignificant	< 0.5% of income	Minor non-compliance	Minor injuries	Minor local impact only.	Low level local adverse publicity - sniping
2 Minor	0.5%-2% of income	Numerous minor non-compliances	Injuries/illness requiring medical treatment - temporary impairment- localised.	Significant impact on local Faculty/Directorate operation.	Medium level local adverse publicity. Critical article. Online public criticism
3 Moderate	2%-5% of income	Major non-compliance. Potential legal challenge/ enforcement action e.g. Improvement Notice.	Injuries or illness requiring hospital admission.	Significant impact on several Faculty/Directorate operations.	National adverse publicity, damage to status of particular Faculties/parts of the University. Threats of public protest.
4 Major	5%-10% of income	Numerous non-compliances. Serious enforcement action e.g. Prohibition Notice	Injury or illness resulting in permanent disability. Multiple injury cases.	Severe disruption to business-as-usual across the University	Widespread adverse publicity, severe damage to status. Ministerial concern/political repercussions. Widespread and ongoing media coverage.
5 Severe – extremely significant	> 10% of income	Widespread non-compliance. Multiple legal challenges. Multiple enforcement actions.	Fatality. Multiple serious injury cases	Significantly threatens business-as-usual across the University.	High degree of national & International adverse publicity, loss of status. Loss of confidence by govt/regulators.

**Likelihood**

1 Rare - lower than 10%
2 Unlikely – lower than 25%
3 Possible – lower than 50%
4 Likely – lower than 75%
5 Very Likely – above 75%

Risk Mitigation	
<b>Tolerate</b>	Accept the risk by keeping activities unchanged when exposure is tolerable, control is impossible or the cost of control exceeds potential benefit. May be supplemented by contingency planning for handling the potential impact. Whether a risk can be tolerated is a key management decision.
<b>Treat</b>	Adjust relevant activities.
<b>Transfer</b>	Share the risk by transferring it using insurance or paying a third party to take the risk.
<b>Terminate</b>	Avoid or cancel the activities that give rise to the risk.

**PRIVACY RISK IDENTIFICATION**

Risk Description (delete where inapplicable)	Impact (a)	Likelihood (b)	Score (a x b)	Mitigating Actions	Rationale	Revised Rating	Action taken
Inability to exercise rights (including but not limited to privacy rights)							
Inability to access services or opportunities							
Loss of control over the use of personal data							
Discrimination							
Identity theft or fraud							
Financial loss							
Reputational damage							
Physical harm							
Loss of confidentiality							
Loss of availability							
Re-identification of pseudonymised data							
Any other significant economic or social disadvantage							
Other							

Part C: DPIA Reporting Form

<b>Date:</b>	9 July 2020	<b>Time:</b>	1pm	<b>Location:</b>	Virtual
<b>Present:</b>	Alison Knight (AMIK) (Chair), Mark Watts (MRW), , Gary Wills (GW), Kelly Davidson (KD), Isobel Stark (IS), Brian Pickering (BP), Heather Smith (HS)				
<b>Apologies:</b>	Silke Roth (SR), Barbara Halliday (BH), Claire Harris (CH)				
<b>DPIA reviewed:</b>	Learner Engagement within Peer-Supported Environments mediated by Digital Technologies: PeerWise and MOOC Platforms				

**Brief Background of DPIA**

See Part A above for background.

DPIA Panel Conclusion:

- **No residual medium or high risks identified or recommendations (including mitigations) proposed by DPIA Panel. Therefore following table intentionally left blank.** Authorised: Dr Alison Knight (Head of Research Integrity & Governance, University of Southampton)

	Risks Identified	Recommendations (including mitigation)	Integration Action	Integration Action by	Due Date	Outcomes	Escalation (if any)
1.							
2.							
3.							

**If Data Protection Officer Advice provided:**

Summary of advice:	Advice accepted/overruled by:	Reasons for overruling:

**Unmitigated high risk/s**

If all or some of the recommendations/actions recommended by the DPIA panel (and, if also obtained and if different from the DPIA Panel's, the recommendations/actions of the Data Protection Officer) are not taken into account then the relevant decision maker not accepting such recommendations/actions must be identified and they must sign below to acknowledge that they understand the unmitigated high risks involved in not taking action, and accept any financial implications of not compliance with GDPR as a consequence including any resulting fines or other punitive consequences that may be imposed for such non-compliance.

	Data Protection Officer	Project Sponsor	*Dean/ Head of School/Other	*UEB (if applicable)
Signature:				
Full name:				
Job Title:				
Date of Approval:				

(Signature is not required if completed electronically).

## APPENDIX 1 – Legitimate Interest Test

This legitimate interest assessment (LIA) template is designed to help you to decide whether or not the legitimate interest basis is likely to apply to your processing. It should be used alongside the ICO's [legitimate interest guidance](#).  
**Not relevant to this DPIA.**

### **Part 1: PURPOSE OF THE TEST**

You need to assess whether there is a legitimate interest behind the processing.

- |  |  |
|--|--|
| <ul style="list-style-type: none"><li>• Why do you want to process the data?</li><li>• What benefit do you expect to get from the processing?</li><li>• Do any third parties benefit from the processing?</li><li>• Are there any wider public benefits to the processing?</li><li>• How important are the benefits that you have identified?</li><li>• What would the impact be if you couldn't go ahead with the processing?</li><li>• Are you complying with any specific data protection rules that apply to your processing (eg profiling requirements, or e-privacy legislation)?</li><li>• Are you complying with other relevant laws?</li><li>• Are you complying with industry guidelines or codes of practice?</li><li>• Are there any other ethical issues with the processing?</li></ul> |  |
|--|--|

B-27

### **Part 2. NECESSITY TEST**

You need to assess whether the processing is necessary for the purpose you have identified.

- |   |  |
|---|--|
| <ul style="list-style-type: none"><li>• Will this processing actually help you achieve your purpose?</li><li>• Is the processing proportionate to that purpose?</li><li>• Can you achieve the same purpose without the processing?</li><li>• Can you achieve the same purpose by processing less data, or by processing the data in another more obvious or less intrusive way?</li></ul> |  |
|---|--|

--

**PART 3: BALANCE TEST**

You need to consider the impact on individuals' interests and rights and freedoms and assess whether this overrides your legitimate interests.

<p><b>Nature of the personal data</b></p> <ul style="list-style-type: none"><li>• Is it special category data or criminal offence data?</li><li>• Is it data which people are likely to consider particularly 'private'?</li><li>• Are you processing children's data or data relating to other vulnerable people?</li><li>• Is the data about people in their personal or professional capacity?</li></ul>
---

--

<p><b>Reasonable expectations</b></p> <ul style="list-style-type: none"><li>• Do you have an existing relationship with the individual?</li><li>• What's the nature of the relationship and how have you used data in the past?</li><li>• Did you collect the data directly from the individual? What did you tell them at the time?</li><li>• If you obtained the data from a third party, what did they tell the individuals about reuse by third parties for other purposes and does this cover you?</li><li>• How long ago did you collect the data? Are there any changes in technology or context since then that would affect expectations?</li><li>• Is your intended purpose and method widely understood?</li><li>• Are you intending to do anything new or innovative?</li><li>• Do you have any evidence about expectations – eg from market research, focus groups or other forms of consultation?</li><li>• Are there any other factors in the particular circumstances that mean they would or would not expect the processing?</li></ul>
--

<b>Likely impact</b>	
<ul style="list-style-type: none"> <li>• What are the possible impacts of the processing on people?</li> <li>• Will individuals lose any control over the use of their personal data?</li> <li>• What is the likelihood and severity of any potential impact?</li> <li>• Are some people likely to object to the processing or find it intrusive?</li> <li>• Would you be happy to explain the processing to individuals?</li> <li>• Can you adopt any safeguards to minimise the impact?</li> </ul>	
Can you offer individuals an opt-out?	Yes / No

29

**MAKING THE DECISION**

This is where you use your answers to Parts 1, 2 and 3 to decide whether or not you can apply the legitimate interest basis.

Can you rely on legitimate interest for this processing?	Yes / No
Do you have any comments to justify your answer? (optional)	



## Data Management Plan

The following is the Data Management Plan (DMP), approved in conjunction with the DPIA documentation (in Appendix B) on the 7th July 2020.

This document is now used by the library team at the University of Southampton as an exemplar for doctoral researchers' training on how to conduct a DMP (<https://library.soton.ac.uk/researchdata/planning>). Some of the text appears also in the body of this thesis.

# Data Management Plan

## About your Research

<b>PhD title:</b>	<b>Clustering Analysis of Learner Engagement within Peer-Supported Environments mediated by Digital Technologies: PeerWise and MOOC Platforms</b>
<b>Student name:</b>	Adriana Wilde
<b>Supervisor(s):</b>	David Millard
<b>Ethics No. (if appropriate)</b>	<b>ERGO/FEPS/55694.A1</b>

## About this plan

<b>Date of plan:</b>	29/06/2020	<b>Frequency of reviews</b>	12m
<b>Date of next review:</b>	June 2021		
<b>Agreed actions to help you implement the plan</b>	<i>Undertake further training for DMP (library RDM pages). Investigate how to transfer responsibility of implementing this plan to my supervisor once I cease to have a Southampton UserID.</i>		
<b>Agreed equipment and/or resources required:</b>	<i>University-provided laptop with a research filestore.</i>		
<b>Further information (as appropriate):</b>			

## Version Table

Version	Changes made	Date
1.0	Creation of this Data Management Plan	26/06/2020
1.1	Section 2: Addition of FutureLearn Reseach Ethics	01/07/2020
2.0	Sections 1,3,7: Following feedback by Michael Whitton from the Research Data Team, I provide further clarification of the nature of the preprocessing, access controls and long term archiving.	05/07/2020

## 1. Project Description:

The rationale of this PhD project lies on the need to understand the patterns of learner engagement using digital technologies in peer-supported environments, such as with PeerWise in the context of face-to-face instruction, and also in the context of massive open online courses (MOOCs). In order to do this, timestamped data of student activity within each of these environment is required, including: questions or comments created, interactions, such as “likes” of peers’ comments or questions, etc. Such data is symbolised as the cloud in Figure 1.1 from my draft thesis, shown below.

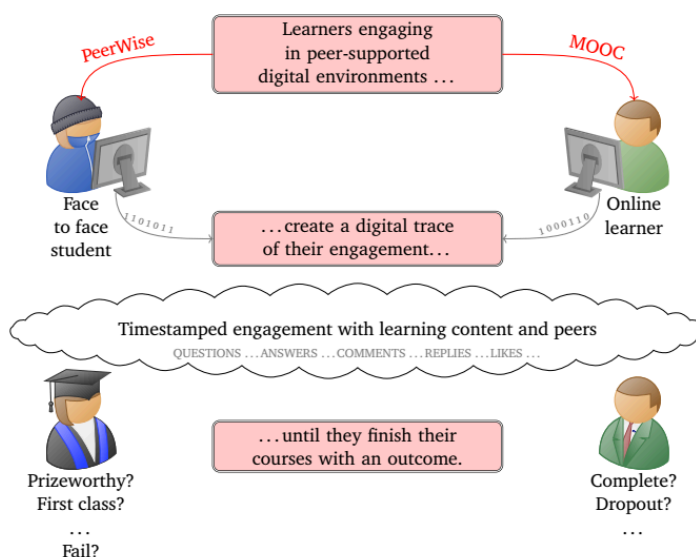


FIGURE 1.1: High-level view of this research in the contexts of face-to-face instruction and MOOCs. Both contexts present learners engaging in a peer-supported digital environment, leaving a data trail of their engagement in the form of timestamped activity (creating content and interacting with their peers through comments and replies, for example). These digital traces are captured in various CSV files in the respective environments, which in turn can be analysed against learner outcomes as shown (attainment in face-to-face instruction, completion in MOOCs).

## 2. What policies will apply to your research?

UoS Code of Conduct for Research (latest version, October 2017).

UoS Policy on the ethical conduct of studies involving human participants (latest version, March 2012).

UoS Research Data Management Policy (University regulations 2019-2020).

UoS Data Protection Policy (latest version, May 2018).

FutureLearn Data Protection Policy <https://www.futurelearn.com/info/terms/data-protection-policy>

FutureLearn Privacy Policy <https://www.futurelearn.com/info/terms/privacy-policy> (last updated on 29 November 2019.)

FutureLearn Research ethics <https://www.futurelearn.com/info/terms/research-ethics-for-futurelearn> (Created on 24th February 2014. Updated on the 21st June 2018).

## 3. What data/research material will you collect or create?

Digital data, of which, some of it is fully secondary data (from MOOCs) and some is partially secondary data (from PeerWise). In addition to these two main sources of secondary data, I will create associated files to the PeerWise courses from preprocessing of assessment and administrative data of the related cohorts of the module “Interaction Design” (COMP2213) at the University of Southampton (UoS). The files are described in section 4, below.

As described in section 1, above, I aim to identify the main patterns of engagement using a specific machine learning technique that involves the grouping of individual data points (clustering). In addition to the description of engagement clusters, the data will be used to predict attainment. In general terms, the research aims to identify whether a specific pattern of engagement is more indicative than others to achieve academic success, i.e. higher marks in the case of face-to-face instruction (using PeerWise data and UoS data) or, in the case of MOOCs, retention and completion, as was shown in Figure 1.1, in section 1, above.

My method for processing the data is described in Figure 1.2, below.

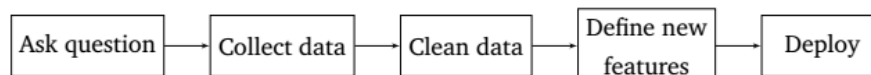


FIGURE 1.2: Data science pipeline applied in this study: experimental setup, data collection, data cleaning, feature extraction, feature selection, classification/clustering, analysis, evaluation of results, insights

In particular, and as part of the “*clean data*” phase of the pipeline shown, I will preprocess my local copy of the original datasets, some of which I already have (e.g. those collected in relation with the modules I taught) and some which will be provided to me by iSolutions once I fulfil the DPIA requirements. Such preprocessing will involve an *automated* part and a *manual* part. The *automated* part, using Python scripts made by me, will aggregate learner activity information and generate a set of features per learner in the following four kinds:

- Features characterising the learner (e.g. in the face-to-face context: their assessment, and organisational details such as membership to peer groups);
- Features capturing temporal information (e.g. when and how often the learner leaves traces of activity);
- Features capturing content production (i.e. engagement with the learning content);
- Features capturing interaction information (i.e. engagement with each other).

Some of these features will be extracted directly from the datasets, other, *derived* features, will be obtained through relatively simple manipulation, and some *high-level* features, will be obtained through more complex manipulation.

One important aspect of the automated data preprocessing involves re-anonymisation of the MOOCs datasets (I will already receive them pseudoanonymised), therefore the profiling will be applied only on fully anonymised data. It is important to stress that no profiling of individuals within a group will lead to their identification, which is why I have added a *manual* part to my preprocessing. This will involve me inspecting the features I generated to ensure any identifiable data (e.g. disclosed by the learners in comments or free-text in surveys) is removed. Therefore, there will be no risk of identifying any individuals. In my dissemination of findings, I will report on clusters of behaviour of anonymous individuals.

#### 4. How will your data/research material be documented and described?

In both cases, the file-naming and file-structure conventions used by the entities who did the data collection in the first place (FutureLearn and PeerWise platforms), will be preserved. For example, the files:

- **MOOCs/portus/1/portus-1\_step-activity.csv** contains step-activity data from the first run of the Portus FutureLearn MOOC.
- **MOOCs/understanding-language/11/understanding-language-11\_comments.csv** contains comments data from the eleventh run of the Understanding Language FutureLearn MOOC.
- **PeerWise/12710\_data/Questions\_12710.csv** contains questions data from the course 12710 in PeerWise (which corresponds to the COMP2213 cohort of 2015/16)

- **PeerWise/14715\_data/Replies\_14715.csv** contains replies data from the course 14715 in PeerWise (which corresponds to the COMP2213 cohort of 2016/17).

In the cases of having creating additional files related to these COMP2213 cohorts (pre-processed, for example from assessment data), I would preserve the file conventions as follows:

- **PeerWise/12710\_data/Groups\_12710.csv** contains data of group formation in the COMP2213 cohort of 2015/16, as in the Student Wiki: <https://secure.ecs.soton.ac.uk/student/wiki/w/COMP2213-1516>.
- **PeerWise/12710\_data/Grades\_12710.csv** contains the grades fields (and only those, associated with the anonymised userID associated with a PeerWise user) extracted from the COMP2213\_data\_grid.xlsm file (which in turn, is associated with the grades awarded to the COMP2213 cohort of 2015/16).

The columns associated to these comma-separated files are listed in full in the following figures (Figures 1.5 and 1.7 for PeerWise, and Figures 1.11 and 1.12 for MOOC data). However, please note that I will not use or require all of the files there described, in particular, I will not use special category data, though in the event any of this kind is disclosed by the participants in free-text (such as comments or survey responses) I will discard it during preprocessing. iSolutions has offered assistance with data minimisation, all the while ensuring that I will be provided with all the data that is actually necessary to achieve my research goals (but no more, as required under the GDPR / DPA 2018.)

The preprocessed data will be documented in a similar way as above described, and described in full within the PhD thesis (most likely in an appendix), in the form of a register document containing the formatted data, indicating the filenames, full title, purpose, date of creation/modification (or whether it was as provided by FutureLearn/PeerWise) as well as its file location.

FIGURE 1.5: Associated Files for the PeerWise datasets (I)

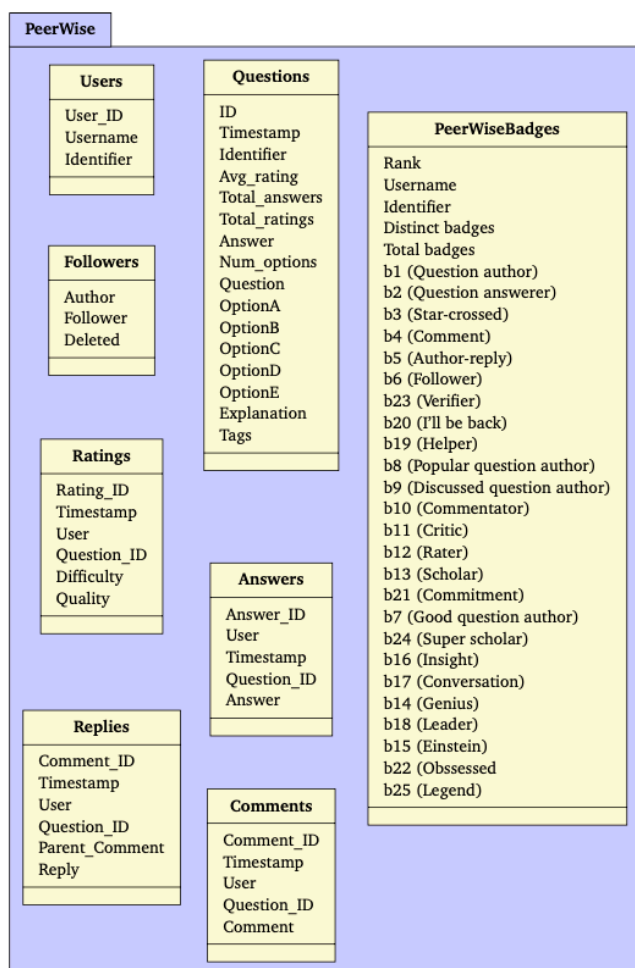


FIGURE 1.7: Associated Files for the PeerWise datasets (II)

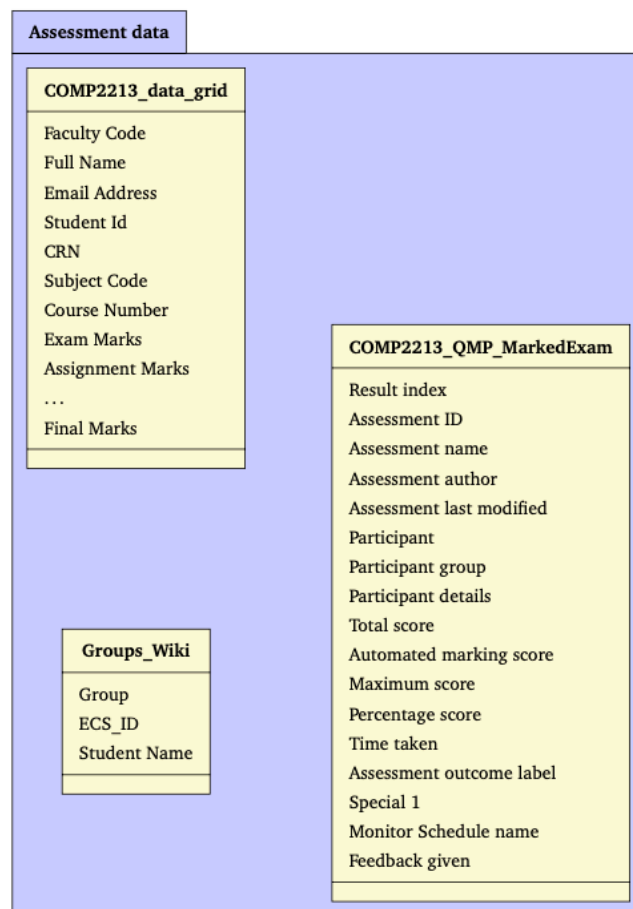
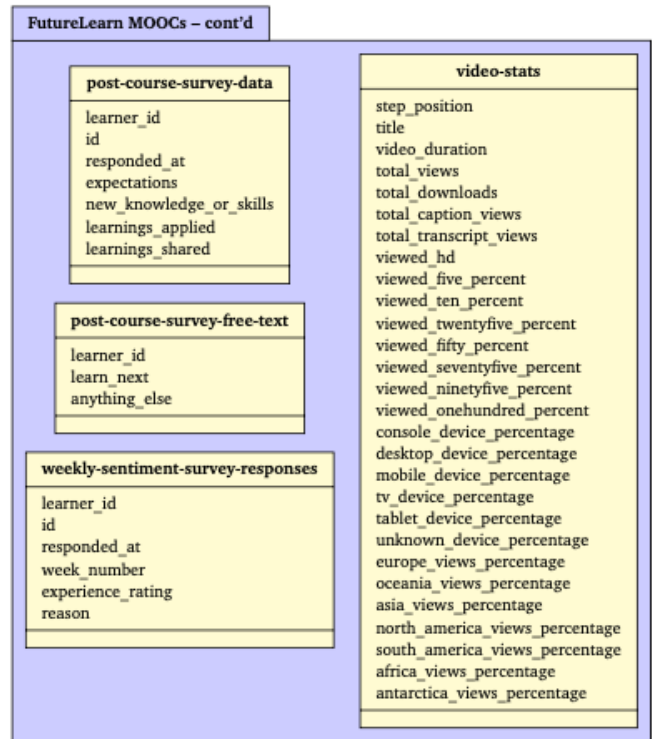


FIGURE 1.11: Associated Files for the FutureLearn MOOCs datasets (part I)



FIGURE 1.12: Associated Files for the FutureLearn MOOCs datasets (part II). Generated by FL only for run 6 of Portus and runs 8 and 9 of Understanding Language)



## 5. How will you deal with any ethical and copyright issues?

In addition to engaging in ethical and research governance processes (this research has been granted the ethics approval numbers ERGO/FEPS/55694 and ERGO/FEPS/55694.A1), I am currently engaging in DPIA processes.

The main ethical issue which could arise here is that the clustering analysis may, for instance, highlight a group of poor performers. As a researcher, it is not my responsibility to fix this problem (in particular, given that an intervention would be impossible to do, as all the courses under analysis have finished long ago). However, I would alert the main stakeholders so that they may want to think about whether current courses should do something specific to investigate these types of students and introduce measures to help. In this context, this would include the current COMP2213 module lead, the course leaders for the Portus MOOC and Understanding Language MOOC in FutureLearn, and FutureLearn themselves, to fulfil the recommendation in their Terms and Conditions (Research Ethics) document:

*“It is also appropriate to provide FutureLearn with a copy of research findings and papers in advance of publication, particularly if these offer any new insight or issues.”*

Additionally, as I do belong to the FutureLearn Academic Network, I can and will communicate any arising ethical or issues with regards to the MOOCs studied which may have wider implications in the FutureLearn community.

## 6. How will your data/research materials be stored, and backed up?

I receive all my secondary data via safesend.soton, use git.soton.ac.uk for version control of the code performing the preprocessing of the data, and use OneDrive for Business/Sharepoint for secure storage of the secondary data as received as well as all the LaTeX files associated with my thesis (including all the bibliography used, in the form of .TeX and .bib files). This means that iSolution will be able to assist with recovering any accidentally deleted files (for up to 90 days after deletion), that the files are encrypted at rest, and that all data is held in secure centers within the UK.

## **7. What are your plans for the long-term preservation of data/research materials supporting your research?**

Clean, preprocessed sections of my generated datasets containing quantitative features may be deposited as a Pure/ePrints data repository and available to the wider research community for reproducibility purposes. Some other sections of the datasets containing qualitative data will either continue to be archived by the University of Southampton (via a research drive managed by my supervisor (as he would inherit this Data Management Plan once I leave the university) or deleted, since all identifiable data will have been permanently removed by me during the cleaning stage of the preprocessing as explained in section 3, above. However, do note that given that the university owns the MOOC data (jointly with FutureLearn), other copies of the original datasets exist and continue to be managed responsibly and in accordance with the policies listed in section 2, above. Should another researcher wish to reproduce/extend my research, they should be able to, as long as they request the original data to the satisfaction of the university's requirements for data protection and data management.

## **8. What are your plans for sharing the data/research materials after the submission of your thesis?**

Findings on the group behaviours emerging from the clustering analyses will be disseminated in relevant venues: in addition to the FutureLearn Academic Network quarterly meetings, in targeted publications (Computers and Education, British Journal of Educational Technologies, Learning @ Scale conference, and the International Conference on Learning Analytics and Knowledge, LAK). In all cases, data for publication will not have two or more direct identifiers, so the pseudonymisation process will not be deconstructed by other researchers resulting in deanonymisation of the published data.

The University of Southampton Library has developed this Doctoral Research Data Management Plan and guidance notes based on material adapted from the Australian National Data Service, Sheffield Hallam University, the Open University and the universities of Bath and Newcastle.



# Explanatory Notes

## What is data?

The term data can be misleading as, in this context, it does not mean Big Data, electronic data or spreadsheets. It means, as the plural of the Latin datum, pieces of information whatever format they are in. In other words, your research materials which you use to answer your research questions and draw your conclusions.

For an historian, these may chiefly be a bibliography or primary and secondary sources and research notes based on those sources, with some additional working copy images of archival material. For a medic, they could be slides of tissue samples, experimental results and patient histories.

## What are data management plans?

A data management plan (DMP) is a document that describes:

- What data/research materials will be created
- What policies will apply to the data/research materials
- Who will own and have access to the data/research materials
- What data management practices will be used
- What facilities and equipment will be required
- Who will be responsible for each of these activities?

## What do I do with this plan?

You should discuss your plan with your supervisory team and it should be uploaded into PGRTracker as part of your progression review documentation. A DMP is a living document so you should revisit it as often as you feel is necessary but at least by every progression review to make sure it is still relevant. Any training or equipment needs which are highlighted in the DMP should be fed into your regular Academic Needs Analysis.

## Why do I need a data management plan?

The carrot: improvements to efficiency, protection, quality and exposure.

Data management in some form is an unavoidable consequence of working with data. Typically data management is done at the last minute and using the first method that comes to mind. This approach is usually time-consuming and error-prone. Taking time at the start of a research project to put in place robust, easy-to-use data management procedures will usually pay off several times over in the later stages of the project. Inadequate data management can also lead to catastrophes like the loss of data or the violation of people's privacy.



The stick: basic data management is required by the University as part of its Data Management Policy, <http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html>, and also by many of the major funders of PhD studentships.

## What does a data management plan need to cover?

The following list of topics can be treated as a check-list:

Backups	This is probably the single most important item on this list. You must have a credible backup strategy of regular backups, and of course you must then follow it. Consider including an off-site backup so that your data will not be lost if your building burns down. Consider an automated backup process.
Survey of existing data	What existing data will need to be managed?
Data to be created	What data will your project create?
Data owners & stakeholders	Who will own the data created, and who would be interested in it?
File formats	What file formats will you use for your data?
Metadata	What metadata will you keep? What format or standard will you follow?
Access and security	Who will have access to your data? If the data is sensitive, how will you protect it from unauthorised access?
Data organisation	How will you name your data files? How will you organise your data into folders? How will you manage transfers and synchronisation of data between different machines? How will you manage collaborative writing with your colleagues? How will you keep track of the different versions of your data files and documents?
Storage	Where will your data be stored? Who will pay for the hardware? Who will manage it?
Bibliography management	What bibliography management tools will you use? How will you share references with the other members of your group/supervisor?
Data sharing, publishing and archiving	What data will you share with others? What license will you apply?
Destruction	What data will you destroy? When? How?
Responsibilities	Who will be responsible for each of the items in this plan?
Anything else	Don't restrict yourself to the items above. Stop and think. What is missing from this list? If you think of something, please let us know so that we can update this information.

# Detailed Guidance Notes

Also see the **Data Plan for your PhD** webpages: <http://library.soton.ac.uk/researchdata/phd>

For additional support, email [researchdata@soton.ac.uk](mailto:researchdata@soton.ac.uk)

## 1. Project Description

Provide two or three sentences summarising your project's research questions and data needs.

## 2. What policies will apply?

University policies that might be relevant to your project are listed below:

- Research Data Management Policy:  
<http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html>
- Open Access Policy: <http://www.calendar.soton.ac.uk/sectionIV/open-access.html>
- Ethics Policies:  
[https://www.southampton.ac.uk/about/governance/regulations-policies-guidelines.page#research\\_%26amp%3B\\_enterprise\\_policies](https://www.southampton.ac.uk/about/governance/regulations-policies-guidelines.page#research_%26amp%3B_enterprise_policies)  
<https://intranet.soton.ac.uk/sites/researcherportal/Lists/Services1/testing.aspx?ID=285&RootFolder=%2A>

### Working with personal or sensitive data?

If humans are involved in your research, you will need to take measures to comply with the Data Protection Act 2018 and GDPR. Consult the University's guidance to find out what this will mean in practice for your research:

<https://www.southampton.ac.uk/legalservices/policy-and-guidance.page>

## 3. What data will you collect or create?

- What physical data will you study? (e.g. artefacts, samples, paper archives, etc.)  
And what digital data will be derived from these? (e.g. field-notes, images, measurements, spreadsheets, survey data, etc.)
- How will the data be collected? Is it gathered from experiments? From the literature? What instruments?  
How about observations or photos?
- Will you be using secondary data?
- Could the data be considered personal, sensitive or commercial data?
- Describe the methods/standards for data creation. What quality assurance processes will you adopt (e.g. calibration, data entry validation, representation with controlled vocabularies)
- What file formats and software will you use? Do your chosen formats and software enable sharing and long-term sustainability of data, such as open standards and open source software?
- Consider how many individual files you expect to make, anticipated file sizes, and total storage volume.
- Frequency of new data - how often will you get new data and over what time period?  
Continuously or just from discrete experiments? How many experiments per week? How will this change over time?

Examples:

*I record interviews with subjects using a digital audio recorder, then transcribe them into text.*

*I test my catalyst under a number of conditions, then submit samples of the products to analysis facilities.*

*I generate data using model code that I've written, then process it in various ways to produce visualisations.*

*I combine existing data from a number of sources [e.g....] and reanalyse them to derive new conclusions.*

*All of my data, part from my literature search, will come from a single 3-month field trip to various archives in France in my second year.*

*I expect to run two or three experiments each week through my second year and much of my third year – about 100 in total*

#### **4. How will your data be documented and described**

Think about what contextual information is required to make the data understandable to others (and yourself in three years' time!).

- Has a file naming convention and directory structure been agreed? (e.g. date created/date amended/version no.)
- What information on the data collection methods and context (documentation and 'metadata') will be recorded for each data type/set?
- Where will the metadata for each data type/set be located? (e.g. within the data file and/or as separate metadata text document, and/or in method chapter/appendices in the thesis)
- How will you tell different versions of the data do you create apart? For example, versions of data files Do you update or add data to existing files?
- Describe the system to name and structure any electronic files. Are there any set or recommended standards in your discipline?

Examples:

*I use the structure <archive collection>/<mss no> for transcripts, notes on documents and working copy images. Filenames are suffixed with transcript, notes or page nos as appropriate.*

*Each filename starts with the date on which the data was collected using the format YYYYMMDD. As I survey new cohorts, data is appended to the dataset and saved as a new file.*

*There is only ever one version of each data file — new experiments create new data, which is stored in a new set of files with machine generated filenames. I keep a register of filenames and the experiments they relate to.*

*Each time I run a new version of my model, intermediate files are written over, but the final results are saved as a new file.*

*Weekly check that files on the R: drive are still usable.*

*Working data is backed up on the UoS Research Filestore. I will make sure I copy the latest versions of my working files there each day.*

*I regularly scan my paper-notebook and store digital copies on the University storage.*

## 5. How will you deal with any ethical and copyright issues?

- Who else has a right to see or use this data, even before you share it? If your data is personal, sensitive or commercial how will you share safely, including plans to anonymise your data?
- Do you need to anonymise data during research or when preparing for sharing, and how will you do this?
- Have you established who owns the copyright in your data?
- If you are re-using someone else's data, are there any restrictions on their re-use?
- Could the data be considered high value and/or vulnerable? For example, is your data likely to attract "hacktivists"? How could this be mitigated?
- How will you destroy any personal, sensitive or commercial data identified above?

Examples:

- *I will share my data with my research group/supervisor using a shared folder Due to the sensitive nature of my data I will encrypt my data and send via Dropoff ([dropoff.soton.ac.uk](mailto:dropoff.soton.ac.uk)) to my collaborators*
- *My data will be pseudo-anonymised prior to sharing, with files encrypted.*
- *My data is of high value and may be subject to commercial sabotage, I will check for advice in the Information Security Best Practice: <https://intranet.soton.ac.uk/sites/gdpr/Pages/Information-security-best-practice.aspx> and contact Information Security team in iSolutions for guidance.*
- *My paper based notes from interviews will be shredded using confidential waste. My electronic files will be overwritten multiple times using specialist software, for example Eraser <https://eraser.heidi.ie/>*

## 6. How will your data be stored and backed up?

- Do you know the backup procedures of the storage space?
- Quantity of data (Megabytes, Gigabytes, Terabytes, other forms of storage)
- Where will the data be stored? For electronic data there should be 3 places, University storage should be one of the locations.
- If keeping your own copy of the data are there security considerations, e.g. encrypted flash drive? How will you know which is the master copy?
- How much have you got so far? Try to estimate how this will grow for the rest of the project
- Describe the regime for backing up the data.
- Describe the procedure to be used to ensure files can be restored from the backups.

Example:

*Each experiment produces about 50MB of data, so over the course of my PhD I expect this to add up to about 5GB, plus two drawers of a standard filing cabinet*

*My primary copy of my bibliography is on my laptop. I make weekly back-ups of it to my University filestore H: drive every Friday afternoon*

## 7. What are the plans for the long-term preservation of data supporting your research?

- What data/research material should be kept beyond the end of the project? Refer to any ethics approval documentation if appropriate.
- What data/research material should be destroyed for contractual, legal or other purposes?
- How long will you preserve your data for?
- Where will you preserve your data? In the UK Data Archive? In the UoS Institutional Repository?
- How will you prepare and document the data for preservation?
- What file formats can you export to for long term preservation?

Examples:

*I am responsible for archiving data, and the archive service will maintain it for a minimum of 10 years as per the University RDM Policy.*

*The data is part of a larger project and will be archived with the project; my supervisor will deal with this.*

*All data, both raw and processed will be retained. Spreadsheets will be saved as csv files.*

*Only simulation code and input parameters will be kept.*

*Transcripts of all interviews, but not recordings. Personal data and anonymization key will need to be destroyed securely at the end of the project.*

## 8. What are your plans for data sharing after submission of your thesis?

- Will any of the digital data supporting the thesis (e.g. organised project archive folders with images, drawings, spreadsheets, databases, etc.) be made available to others via a repository?
- Are there funding body/institutional requirements for the re-use of, or open-access to, the data?
- What are your supervisor's thoughts on sharing 'their' research data, if on a project team?
- With whom will you share your data and under what conditions? Should anybody be able to download the data, or is there a need for access restrictions (e.g. an embargo period, or making data available on request only)?
- Who, if any, are the anticipated future users of any digital data/resources from the research, e.g. yourself, project partners, future students, peer researchers, the public?
- Where will the data be archived?
- Who will create and maintain the archive of data?

Examples

*Tables for household income and relative market prices of goods in my thesis will be made available as spreadsheets. My bibliography will be made available as a csv file so it can be reused by other scholars. These will be made publicly available in the institutional repository and linked to my thesis. The working copy images of archival material cannot be shared due to copyright restrictions by the various archives, however I will upload the full transcripts of those documents which I quote extensively in my thesis if allowed by the relevant archives.*

*All my experimental data will be made available on the institutional repository, accompanied by a readme file describing the data and the data linked back to the relevant part of my thesis. The data will only be made available after a three year embargo period as I plan to publish further articles from my thesis.*



## Additional details for MOOCs datasets

The tables in the following pages provide additional information about the datasets used in this Thesis. In particular, Tables [D.1](#), [D.2](#) and [D.3](#) show the shapes of the files in the datasets as received from iSolutions in May 2020, following Ethical Approval in March (detailed in Appendix A). In particular, the number of lines and columns to each file. The names of the columns are listed in Figures [3.3](#) and [3.4](#), in Chapter 3.

Through the inspection of those datasets it became evident that the anonymisation process by the University had unfortunately rendered them unusable for this research, as the `learner_id` column had been removed from most files (all, except `question-response.csv`) and therefore it was impossible to construct a learner-centred feature vector from this data. This led me to engage in the Data Protection Assessment process detailed in Appendix [B](#) to procure data that could be used to generate learner-centred features, whilst thoughtfully planning how to protect the individuals who generated it.

TABLE D.1: Files in the Portus MOOC dataset

Run	CSV file name	rows	columns
1	enrolments	7773	13
	step-activity	281159	5
	comments	20253	12
	question-response	133749	10†
	peer-review-assignments	265	9
	peer-review-reviews	659	9
2	enrolments	8920	13
	step-activity	213537	5
	comments	18846	12
	question-response	100842	9
	peer-review-assignments	356	9
	peer-review-reviews	681	9
3	enrolments	3252	13
	step-activity	58559	5
	comments	3566	12
	question-response	26840	9
	peer-review-assignments	89	9
	peer-review-reviews	109	9
4	enrolments	5172	13
	step-activity	94904	5
	comments	13929	12
	question-response	47329	9
5	enrolments	4266	13
	step-activity	84356	5
	comments	12465	12
	question-response	43477	9
6	enrolments	1286	13
	step-activity	41964	5
	comments	5010	12
	question-response	21821	9
	weekly-sentiment-survey-responses	9	5
	video-stats	73	28
	archetype-survey-responses	157	3
leaving-survey-responses	85	7	

† Note that `question-response.csv` has one additional column in the first run of this course only, because `learner_id` appears only in that file. It was stripped from all the other files in both datasets.

TABLE D.2: Files in the understanding-language MOOC dataset (runs 1..8)

Run	CSV file name	rows	columns
1	enrolments	58782	13
	step-activity	467333	5
	comments	145425	12
2	enrolments	41913	13
	step-activity	317324	5
	comments	86139	12
3	enrolments	44284	13
	step-activity	228623	5
	comments	58285	12
4	enrolments	25591	13
	step-activity	197266	5
	comments	50332	12
5	enrolments	19873	13
	step-activity	127584	5
	comments	37637	12
6	enrolments	10279	13
	step-activity	73588	5
	comments	18616	12
7	enrolments	12900	13
	step-activity	115204	5
	comments	24941	12
	archetype-survey-responses	1586	3
	leaving-survey-responses	185	7
8	enrolments	6034	13
	step-activity	41149	5
	comments	9307	12
	archetype-survey-responses	607	3
	leaving-survey-responses	128	7
	weekly-sentiment-survey-responses	140	5
	video-stats	32	28



TABLE D.3: Files in the understanding-language MOOC dataset (runs 9..11)

Run	CSV file name	rows	columns
9	enrolments	8311	13
	step-activity	74832	5
	comments	16368	12
	archetype-survey-responses	946	3
	leaving-survey-responses	137	7
	weekly-sentiment-survey-responses	232	5
	video-stats	35	28
	post-course-survey-data	164	6
	post-course-survey-free-text	64	2
10	enrolments	5096	13
	step-activity	41564	5
	comments	8469	12
	archetype-survey-responses	503	3
	leaving-survey-responses	78	7
	weekly-sentiment-survey-responses	116	5
	video-stats	35	28
	post-course-survey-data	126	6
	post-course-survey-free-text	49	2
11	enrolments	7832	13
	step-activity	61110	5
	comments	13229	12
	archetype-survey-responses	798	3
	leaving-survey-responses	128	7
	weekly-sentiment-survey-responses	205	5
	video-stats	35	28
	post-course-survey-data	166	6
	post-course-survey-free-text	64	2

TABLE D.4: Summary table of entries per file in the Portus MOOC dataset per run

file	Run					
	1	2	3	4	5	6
enrolments	7773	8920	3252	5172	4266	1286
step-activity	281159	213537	58559	94904	84356	41964
comments	20253	18846	3566	13929	12465	5010
question-response	133749	100842	26840	47329	43477	21821
peer-review-assignments	265	356	89			
peer-review-reviews	659	681	109			
weekly-sentiment-survey-responses						9
video-stats						73
archetype-survey-responses						157
leaving-survey-responses						85



TABLE D.6: Dates in the portus MOOC dataset

Run	First activity	Start course	End course	Last activity
1	11 Apr 2014	19 May 2014	30 Jun 2014	12 Jul 2014
2	2 Nov 2014	26 Jan 2015	9 Mar 2015	21 Mar 2015
3	24 Mar 2015	15 Jun 2015	27 Jul 2015	8 Aug 2015
4	18 Mar 2016	13 Jun 2016	25 Jul 2016	6 Aug 2016
5	21 Nov 2016	30 Jan 2017	13 Mar 2017	8 Apr 2017
6	2 Feb 2018	26 Feb 2018	9 Apr 2018	5 May 2018

TABLE D.7: Dates in the understanding-language MOOC dataset

Run	First activity	Start course	End course	Last activity
1	23 Oct 2014	17 Nov 2014	15 Dec 2014	27 Dec 2014
2	9 Feb 2015	20 Apr 2015	18 May 2015	30 May 2015
3	9 Sep 2015	19 Oct 2015	16 Nov 2015	28 Nov 2015
4	9 Feb 2016	4 Apr 2016	2 May 2016	14 May 2016
5	14 Sep 2016	17 Oct 2016	21 Nov 2016	10 Dec 2016
6	22 Feb 2017	24 Apr 2017	29 May 2017	24 Jun 2017
7	1 Nov 2017	8 Jan 2018	5 Feb 2018	24 Feb 2018
8	11 May 2018	11 Jun 2018	9 Jul 2018	4 Aug 2018
9	18 Sep 2018	22 Oct 2018	19 Nov 2018	15 Dec 2018
10	2 Apr 2019	29 Apr 2019	27 May 2019	22 Jun 2019
11	5 Sep 2019	21 Oct 2019	18 Nov 2019	14 Dec 2019

## Synthetic data for the example in Figure 6.1 under PeerWise

Table E.1 shows the shape of the synthetic dataset for PeerWise containing engagement data for hypothetical learners  $s_1$ ,  $s_2$  and  $s_3$ , created to test the feature engineering process inspired by the platform-agnostic model of learner engagement in peer-supported learning environments that was defined in Chapter 4.

Sample contents to these files are shown in Tables E.2 to E.7.

TABLE E.1: Files in the test dataset for PeerWise

File name	rows	columns
Users_test	3	3
Questions_test	7	16
Comments_test	5	5
Replies_test	6	6
Followers_test	1	3
Ratings_test	2	6
Badges_test	3	26
Answers_test	7	5
Groups_test	3	3
Grades_test	3	27

TABLE E.2: Contents of `Users_test.csv` for the example in Figure 6.1.

User_ID	Username	Identifier
$s_1$	$s_1$	$s_1$
$s_2$	$s_2$	$s_2$
$s_3$	$s_3$	$s_3$

TABLE E.3: Contents of `Questions_test.csv` for the example in Figure 6.1. (some fields omitted)

ID	...	Identifier	...	Explanation	...
$q_{1,1}$	...	$s_1$	...	question $q_{1,1}$	...
$q_{1,2}$	...	$s_1$	...	question $q_{1,2}$	...
$q_{2,1}$	...	$s_2$	...	question $q_{2,1}$	...
$q_{2,2}$	...	$s_2$	...	question $q_{2,2}$	...
$q_{2,3}$	...	$s_2$	...	question $q_{2,3}$	...
$q_{2,4}$	...	$s_2$	...	question $q_{2,4}$	...
$q_{3,1}$	...	$s_3$	...	question $q_{3,1}$	...

TABLE E.4: Contents of `Comments_test.csv` for the example in Figure 6.2.

Comment_ID	Timestamp	User	Question_ID	Comment
$c_{1,1,1}^2$	...	$s_2$	$q_{1,1}$	This is comment $c_{1,1,1}^2$ made by $s_2$ on $q_{1,1}$
$c_{1,1,2}^2$	...	$s_2$	$q_{1,1}$	This is comment $c_{1,1,2}^2$ made by $s_2$ on $q_{1,1}$
$c_{2,3,1}^3$	...	$s_3$	$q_{2,3}$	This is comment $c_{2,3,1}^3$ made by $s_3$ on $q_{2,3}$
$c_{2,4,1}^1$	...	$s_1$	$q_{2,4}$	This is comment $c_{2,4,1}^1$ made by $s_1$ on $q_{3,1}$
$c_{3,1,1}^3$	...	$s_1$	$q_{3,1}$	This is comment $c_{3,1,1}^3$ made by $s_1$ on $q_{3,1}$

TABLE E.5: Contents of Replies\_test.csv for the example in Figure 6.1.

Comment_ID	Timestamp	User	Question_ID	Parent_Comment	Reply
$r_{1,1,1,1}^1$	...	$s_1$	$q_{1,1}$	$c_{1,1,1}^2$	yes
$r_{1,1,1,2}^2$	...	$s_2$	$q_{1,1}$	$c_{1,1,1}^2$	yes
$r_{2,3,1,1}^2$	...	$s_2$	$q_{2,3}$	$c_{2,3,1}^3$	yes
$r_{2,4,1,1}^3$	...	$s_3$	$q_{2,4}$	$c_{2,4,1}^1$	yes
$r_{3,1,1,1}^3$	...	$s_3$	$q_{3,1}$	$c_{3,1,1}^3$	yes
$r_{1,1,2,1}^2$	...	$s_1$	$q_{1,1}$	$c_{1,1,2}^2$	no

TABLE E.6: Contents of Ratings\_test.csv for the example in Figure 6.2.

Rating_ID	Timestamp	User	Question_ID	Difficulty	Quality
rat111	...	$s_2$	$q_{1,1}$	0	4
rat112	...	$s_3$	$q_{1,2}$	0	4

TABLE E.7: Contents of Answers\_test.csv for the example in Figure 6.2.

Answer_ID	User	Timestamp	Question_ID	Answer
$a_{1,1}$	$s_1$	...	$q_{2,1}$	B
$a_{1,2}$	$s_1$	...	$q_{2,2}$	B
$a_{1,3}$	$s_1$	...	$q_{2,3}$	B
$a_{2,1}$	$s_2$	...	$q_{1,1}$	B
$a_{2,2}$	$s_2$	...	$q_{1,1}$	B
$a_{3,1}$	$s_3$	...	$q_{2,4}$	B
$a_{3,2}$	$s_3$	...	$q_{1,1}$	B

## Coursework specification for participation in PeerWise (COMP2213)

The following pages show the coursework specification<sup>1</sup> for the Interaction Design (COMP2213) cohort of 2015/16, mentioned in Section 6.1.

Immediately after this document, here are also the instructions for registering in PeerWise<sup>2</sup> that the students were instructed to follow.

The cohort of 2016/17 were given optional joining instructions (very similar to those in pages F-5 and F-6, but with course ID 14715 instead of 12710 and no rewards for participation).

---

<sup>1</sup>Available at: <https://secure.ecs.soton.ac.uk/noteswiki/images/COMP2213-1516-CW2.pdf> (only accessible through the departmental intranet. Retrieved 27 July 2020).

<sup>2</sup>Available at: <https://secure.ecs.soton.ac.uk/noteswiki/images/COMP2213-1516-PeerWise.pdf> (*ibid.*)

## Assignment Instructions (COMP2213)

<b>Module:</b>	Interaction Design	<b>Lecturers:</b>	agw106 ss33g15
<b>Assignment:</b>	Comprehension (PeerWise)	<b>Weighting:</b>	10%
<b>Deadlines:</b>	15 <sup>th</sup> March 2016, 13:00 (midway point) 26 <sup>th</sup> April 2016, 14:00 (final questions) 5 <sup>th</sup> May 2016, 14:00 (reflective essay)	<b>Feedback:</b> By email within 2 weeks	<b>Effort:</b> ~15 hours

### About the coursework

Involving students as partners in assessment activities help the development of your expertise and enables a greater understanding of what constitutes high-quality work.

For this coursework, you are required to formulate multiple choice questions (MCQ) on topics in Interaction Design to aid your exam revision. This would enable you to have a good understanding of the examiners' likely frame of mind when producing questions on the examinable content ([McMillan & Weyers, 2011](#)). In addition, you will be supporting each other's learning by answering questions formulated your peers, and offering your feedback, demonstrating your comprehension of the materials covered in this module. To support this work, the online platform PeerWise (<https://peerwise.cs.auckland.ac.nz/>) is used. Instructions for registering can be found [in the NotesWiki](#).

This coursework counts for 10% of the credit for this module, which is a 150 nominal hour module. You should therefore devote **approximately** 15 hours in total to complete this coursework, including the background reading necessary to formulate and answer questions in PeerWise.

### Submission

This **group** coursework has two aspects: Contribution (5%) and Reflection (5%).

The **contribution** will be monitored within PeerWise, specifically at two points:

- by 15<sup>th</sup> March (midway point) and
- by 26<sup>th</sup> April (last date for any questions).

Each person in your group must create a total of 4 questions and answer a minimum of 4 questions on PeerWise (submitted by students in other groups), following the guidelines offered in class to making good MCQ providing constructive feedback on the questions they answer. It is important that the group facilitates peer support to ensure all members contribute. However, we acknowledge that occasionally work is not equally distributed in groups, and for this reason you will declare the actual distribution using [this](#) form.



Note that no traditional “submission” is required for the contribution component as the examiners can access both the statistics offered by the PeerWise system and the individual engagement in the system (questions/answers). Please also note that your contribution is not anonymous (even if it seems so), therefore please be judicious in your remarks at all times. There will be zero tolerance for bullying, harassment and victimisation as per University policy ([http://www.southampton.ac.uk/diversity/policies/dignity\\_at\\_work.page](http://www.southampton.ac.uk/diversity/policies/dignity_at_work.page))

The **reflection** part of the coursework requires for your group to produce a 500-750 word (one A4 page) essay on how the use of PeerWise has facilitated your learning of Interaction Design and whether it has contributed to a greater understanding of the topic. This reflective essay is in the form of a formatted Microsoft Word document, converted to PDF. Any in-text references do not contribute to the word count.

These files (together with the [Coursework Distribution Form](#)) are to be submitted electronically via C-BASS (<http://handin.ecs.soton.ac.uk>) before the 5<sup>th</sup> May deadline. Do not submit hard copies. In each occasion you will need to submit the .pdf file (together with the source file if produced electronically, i.e. typically the .doc file before the .pdf is created). Do not forget to convert your work to PDF format just before submitting.

Note that the standard ECS late penalties apply to the reflective essay, as detailed in the regulations (para. 4.1 of <http://www.calendar.soton.ac.uk/sectionXII/ecs-ug.html>). They are 10% per working day that a piece of work is overdue, up to a maximum of 5 days, after which the mark becomes zero.

## Feedback

You will receive feedback on this coursework via email within 2 weeks of your submission.

## Learning Outcomes

On successful completion of this work you will demonstrate knowledge and understanding of:

- A1. How different disciplines (human factors, cognitive psychology, engineering, graphics design, etc.) influence the design of interactive systems
- A2. How users interact (dialogue) with systems.
- A3. The classification of input/output devices and techniques
- A4. How to design, prototype and evaluate a user interface

You will also be able to:

- B1. Describe the main concepts (conceptual model, metaphors and paradigms) that influence human-computer interaction
- B2. Explain the main theories of cognition and how these are used when designing interactive systems
- B3. Classify the different input/output devices as to their effect on human-computer interaction.
- B4. Describe the process of designing for interaction and why a user centred approach is preferred.

All the while demonstrating ethical and professional values.

## Marking Scheme

Your coursework will be marked out of 10. The marking criteria below will be used:

Criterion	Description	Outcomes	Total
<i>Contribution</i>	The extent to which individuals in the group have contributed	A1, A2, A3, A4.	5 marks
<i>Reflection</i>	The extent to which the Reflective essay evaluates the use of PeerWise, its strengths and weaknesses and transferable skills.	B1, B2, B3, B4	5 marks

Please note that the University regulations regarding academic integrity apply (<http://www.calendar.soton.ac.uk/sectionIV/academic-integrity-regs.html>).

The following descriptors will be used as guidance:

### **Contribution**

- All members of the group created and answered timely at least 4 relevant questions (5 marks)
- Some members did not fully participate, created irrelevant questions, or created all their content after the mid-way point, suggesting late engagement (2-4 marks)
- Very low engagement across the group (0-1 mark)

### **Reflection**

- Concise, considered, insightful reflection on the use of PeerWise. Insightful discussion on the strengths, weaknesses, and insights from the system and evidence of understanding of how these insights may be applied to other areas of study. (4-5 marks)
- Adequate reflection on the use of PeerWise. Adequate discussion of strengths, weaknesses and insights from the system. Some understanding demonstrated of the applicability of the insights to other areas of study. May lack conciseness or not meet the word count limits. (2-3 marks)
- Inadequate reflection on the use of PeerWise, very limited discussion on strengths or weaknesses. May be hard to understand, poorly structured, or present a high number of grammatical errors. (0-1 mark)

## Sourcing reference material

Add as a footnote the list of resources used (but exclude them from the final word count).

Please note that the University regulations regarding academic integrity apply

(<http://www.calendar.soton.ac.uk/sectionIV/academic-integrity-regs.html>)

## Instructions on how to register in PeerWise for COMP2213 so that your participation is correctly recorded.

**Step 1:** Go to: <https://peerwise.cs.auckland.ac.nz/>. You will see the following screen, and will type (and select) "University of Southampton (Southampton, England)." Click Go>>

PeerWise  
Ask | Share | Learn

JOIN NOW  
Get started!  
Follow @peerwise

Welcome to PeerWise

To log in, select your school / institution from the list below

University of Southampton Go »

Just type the first few characters...

**Step 2:** You will be presented with the following page:

PeerWise  
University of Southampton

Welcome to PeerWise

PeerWise supports you and your peers in the creation, sharing, evaluation and discussion of assessment questions relevant to your studies.

**You design the questions**  
Creating a question requires you to reflect on what you are learning in a course. Explaining the answer to your question in your own words helps to reinforce your understanding. *If you teach it, you understand it.*

**See what everyone thinks**  
Attempt questions written by your peers, and see how everyone else has answered. Feedback is immediate, you have access to explanations and you can participate in discussions. *See what others think is important.*

**Learn from your peers**  
Search by quality, difficulty and topic to find questions of interest to you. Follow authors who contribute questions that you like, and request help when you need it. *Help your peers, and let them help you.*

Welcome to PeerWise for the University of Southampton

Already joined? Welcome back...

username:  login »

password:

Forgotten your password? Get a new one  
Forgotten your username? Recover it

Like to join? Please register...

Registration is very simple

PeerWise is simple to use - you can access it anywhere and anytime. **New to PeerWise?** Find out all you need to know.

Follow @peerwise

Click on "Registration" the first time you are presented this page. (In subsequent log ins, just use your UoS username and chosen password for PeerWise)

You will then need to supply a "name" (which will be your UoS username), select a **password** for your PeerWise account, the "Course ID" (which is **12710** – NOT COMP2213), and supply your "identifier", which is your student number. The latter has been pre-loaded into PeerWise for authentication, so ensure you input it carefully.

### Registration

Welcome to PeerWise! Registration is very simple, and consists of the following 4 steps:

- Step 1:** choose a **name**
- Step 2:** choose a password
- Step 3:** enter the "Course ID" for the course you would like to join
- Step 4:** enter your "Identifier" to join the course



### What do I need to know before I start?

Before you start the registration process, you need to know details of the first course that you are going to join. Make sure you know the following **two** things. Your course instructor should have given you this information.



**Course ID** this number identifies the course that you are going to join



**Identifier** this is the information about you that will help your instructor identify you

### I'm ready!

[Begin registration »](#)

## Step-centred analysis of the first run of Understanding Language

Prior to conducting the learner-centred analysis that is the focus of this thesis, I conducted a preliminary activity-centred study, specifically step-centred. Here, the feature extraction process was not guided by the model in Chapter 4. Instead, other features were explored (including temporal features), all extracted from the step-activity file, associated to this MOOC (`understanding-language-1_step-activity.csv`).

This file contains the columns `learner_id`, `step`, `week_number`, `step_number`, `first_visited_at`, `last_completed_at`, as listed in Figure 3.3, and I used it to extract or calculate the following features (with the exception of *Step\_Type*, as explained below):

- **step**: Unique step ID as per each entry on the step-activity file.
- **first\_vis\_day**: Day of the week in which the step was first visited by a learner.
- **first\_vis\_hour**: Hour of the day in which the step was first visited by a learner. The time is recorded in UTC, but in this first run of the MOOC the vast majority of the learners were based in the UK.
- **when\_first(fine)**: In this feature, time of day was divided into slots of three hours each, to record when the step was first visited.
- **when\_first(coarse)**: In this feature, time of day was divided into slots of six hours each.

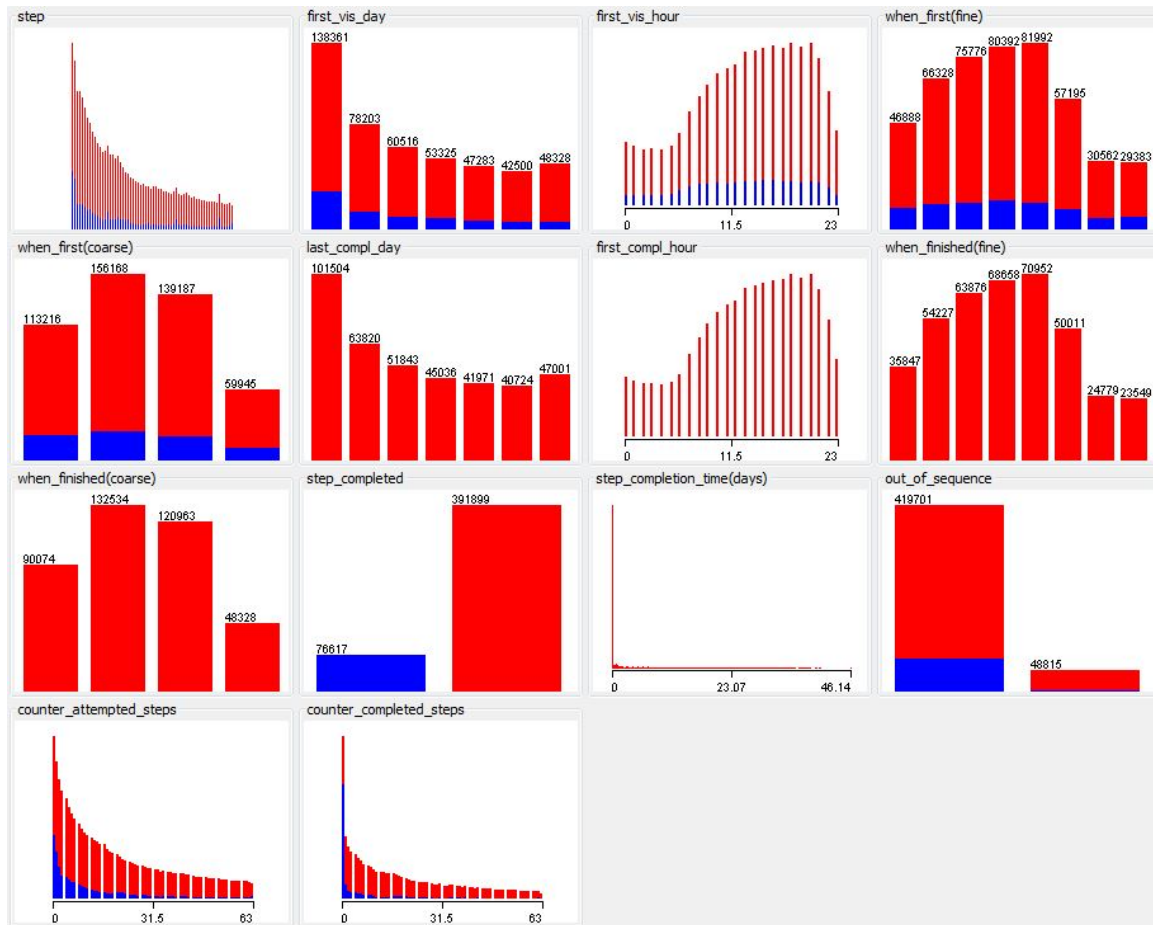


FIGURE G.1: WEKA Explorer visualisation of features extracted from the step-activity file associated to the first run of the Understanding Language MOOC.

- **last\_compl\_day**: Day of the week in which the step was last completed by a learner.
- **last\_compl\_hour**: Hour of the day in which the step was first visited by a learner.
- **when\_finished(fine)**: the 3-hour slot in which the step was last completed.
- **when\_finished(coarse)**: the 6-hour slot in which the step was last completed.
- **step\_completed**: a Boolean to indicate whether the step was completed or not. This is shown in Figure G.1 in red when completed, and blue when not.
- **step\_completion\_time(days)**: time difference (in days) between first visited and last completed.
- **out\_of\_sequence**: a Boolean to indicate whether the step was completed after a step with a higher label.

- 
- **counter\_attempted\_steps**: In this run of the MOOC, there were 64 distinct learning steps, each of them of one the following types: article, audio, discussion, exercise and video (as identifiable in Figure G.2). This counter sums the number of times learners attempted each of the 64 steps.
  - **counter\_completed\_steps**: As above, a counter for each of the 64 steps, only that it focuses on completion.
  - **Step\_Type**: Either article, audio, discussion, exercise or video. This information is not contained in the step-activity file, and was provided separately by FutureLearn.

The variables listed above were used to generate Figures G.2 on the first run of the FutureLearn Understanding Language MOOC. This course ran in November 2014 and had 58,782 learners (as shown in Table D.7). In this run of the MOOC, there were 64 distinct learning steps. I did not have the step type information about any of the other runs of this MOOC, but I was aware that there had been some variations in the learning design over the years, as evidenced in Table 5.2. Therefore, I did not incorporate this feature in my learner-centred feature set used in this Thesis.

Finally, the box-and-whisker plots in Figures G.3 and G.4 show the spread, median and first- and third quartiles of steps visited and completed, respectively (grouped according to their type). The observation from inspecting those visualisations is that the type of step does not seem to have a significant effect on whether the step is completed, and in fact it seems that the later steps will have a lower likelihood to be completed, regardless the type of step, as we can observe the “funnel of participation” effect of MOOC learning, as coined by Clow (2013). An interesting extension to this study would be to confirm this interpretation by studying the type of activity a learner is more likely to complete, should this information become available for all the runs of the MOOCs under study.

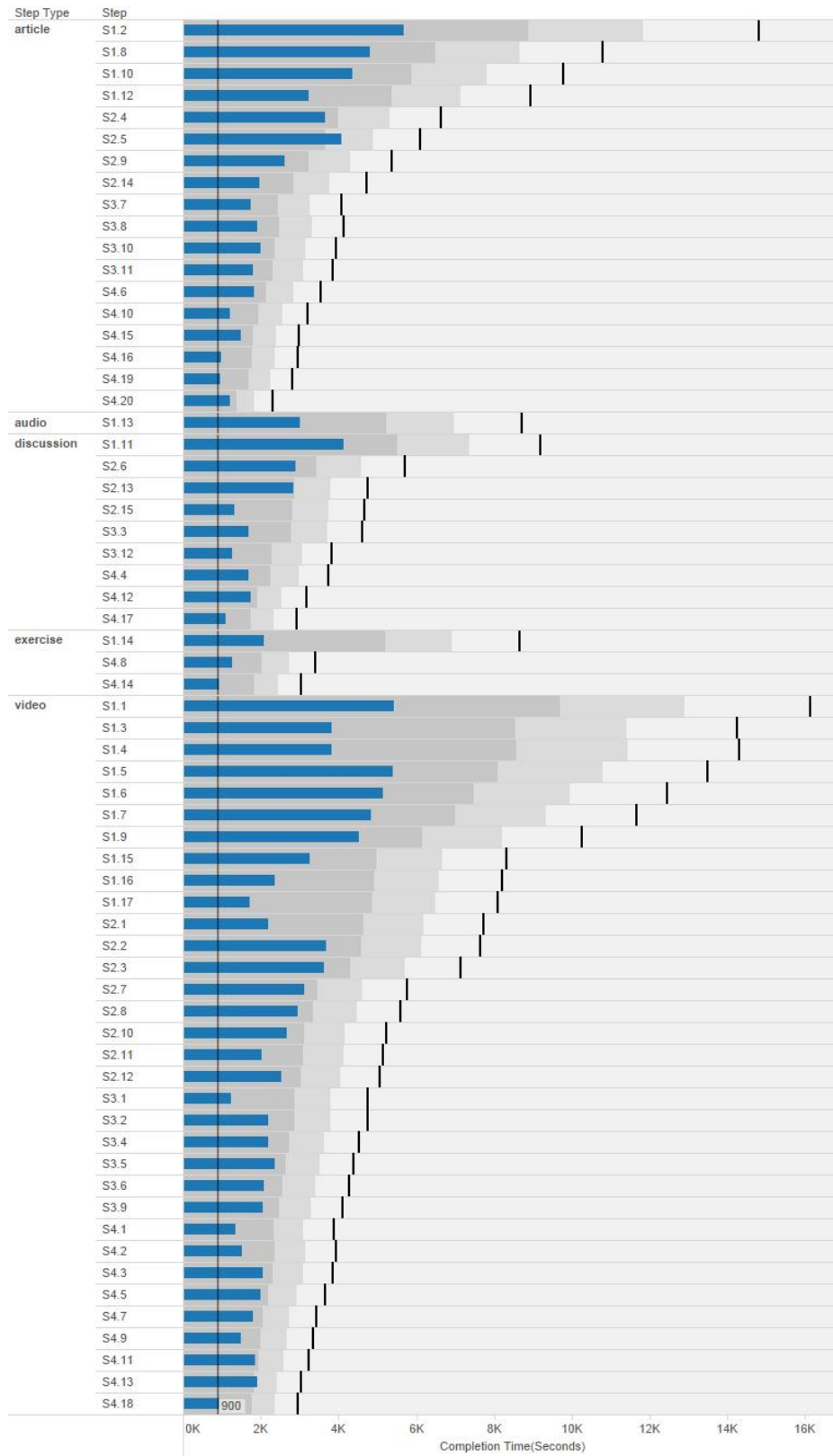


FIGURE G.2: Distinct count of completion time for each step, organised by step type.



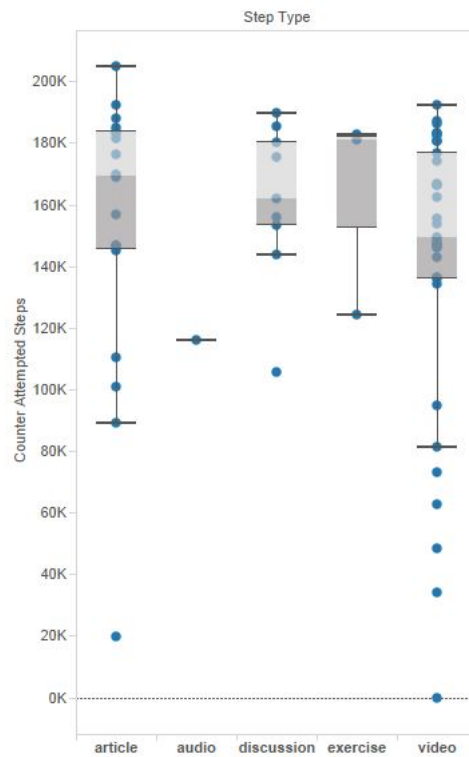


FIGURE G.3: Visited steps in the Understanding Language MOOC (run 1), per step type.

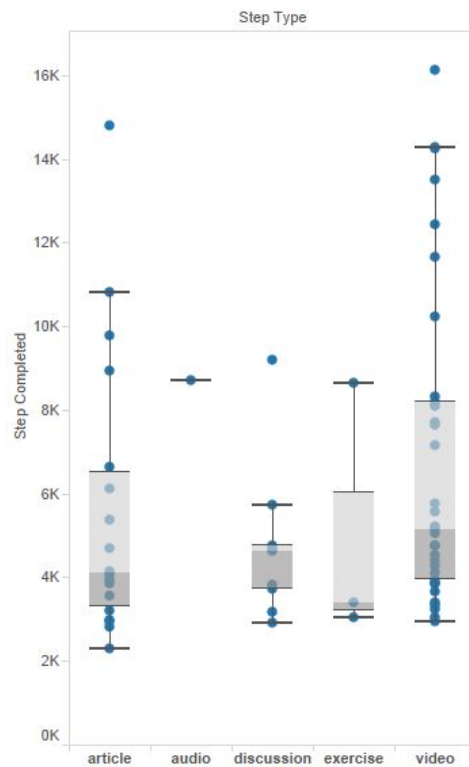


FIGURE G.4: Completed steps in the Understanding Language MOOC (run 1), per step type.

## Clustering on Interval Features in FutureLearn MOOCs

I performed a set of experiments as part of my interest in validating the model presented in Chapter 4, which includes temporal interval features. These were excluded from the main analysis as they fall outside of the main narrative in which I aimed to compare findings across both kinds of peer-supported environments studied and these features are not derivable from the PeerWise dataset available to me during this research. However, as explained, PeerWise keeps logs of interval data (see footnote 2 in subsection 7.5.1) and an interesting extension to my research would look into extracting such features once that additional data is procured.

The features used in this set of experiments are those defined by Allen’s interval algebra, listed in Table 5.6. For illustrative purposes, in this appendix I present the first of such experiments. The dataset used was the first run of Understanding Language with 58,781 enrolled learners, and a total of 467,332 distinct steps, labelled and counted per learner who produced them, as per the interval features above. The resulting pre-processed dataset, `INTV_understanding-language1.arff` with this reduced set of features was subjected under the Expectation Maximisation clusterer in WEKA (EM), with the following distinct four clusters being found by the algorithm:

Cluster 0 **Sequential completer**: This cluster has the highest count of *precede* steps (with a mean±std.dev of  $52.09 \pm 10.32$  steps), and the lowest count of *abandoned* steps. This cluster contains 2,795 learners (10% of the cohort).

Cluster 1 **Early dropout**: This cluster is characterised by having the lowest count of *precede* steps ( $1.38 \pm 2.25$  steps), and the lowest counts in almost all features apart from

*abandoned* (though it was still a low count, with a mean±std.dev of  $2.45\pm 2.25$  steps, suggesting they dropped out too early to even sample more of the MOOC. This cluster contains 15392 learners (55%).

Cluster 2 **Late dropout:** This cluster is characterised by completing steps sequentially before dropping out, with the third largest count of *precede* steps (with  $23.75\pm 16.62$  steps). This cluster contains 6,470 learners (23%).

Cluster 3 **Sampler completer:** This cluster has the highest number of abandoned steps (with a mean of  $8.41\pm 11.96$ ) and the second highest number of *precede* steps (with a mean of  $23.76\pm 16.62$ ). This cluster contains 3,301 learners (12% of the total).

The above interpretation of clusters' semantics is based solely on the information given by the EM algorithm about the centroids of each cluster. This is a common practice in the community, as seen in research by [Bogarín et al. \(2014\)](#) and [Romero, Cerezo, Bogarín, and Sánchez-Santillán \(2016\)](#).

## H.1 Expectation Maximisation clustering with interval features

What follows is the raw output for an experiment on Understanding Language (run 1) using the Expectation Maximisation (EM) clustering algorithm as implemented on WEKA. The clusters found have been interpreted semantically as above, and annotated in highlighting for readability.

```
=== Run information ===
```

```
Scheme:      weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll
              -iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:    Reduced/understanding-language1_-weka.filters.unsupervised.attribute
              .Remove-R1
Instances:   27958
Attributes:  8
              precede
              overlap
              during
              abandoned
              equal
              finish
```

```

                meet
                last
Test mode:     evaluate on training data

```

```
=== Clustering model (full training set) ===
```

```
EM
```

```
==
```

```
Number of clusters selected by cross validation: 4
```

```
Number of iterations performed: 3
```

Attribute	Cluster			
	0 (0.1)	1 (0.57)	2 (0.22)	3 (0.11)
=====				
precede				
mean	52.0902	1.3848	13.3848	23.7573
std. dev.	10.3187	2.2477	7.5328	16.6206
overlap				
mean	4.9381	0.0822	0.8517	3.5268
std. dev.	8.2041	0.3093	1.3195	5.3472
during				
mean	3.5984	0.0206	0.687	1.8
std. dev.	3.6498	0.142	0.9262	2.2045
abandoned				
mean	0.3485	2.449	1.3669	8.4157
std. dev.	0.7656	2.6872	1.2727	11.9664
equal				
mean	0	0	0	0
std. dev.	0	0	0	0
finish				
mean	0.0014	0	0	0
std. dev.	0.0371	0.012	0.012	0.012
meet				
mean	1.4831	0.0001	0.4967	0.8544
std. dev.	2.7121	0.0091	0.8524	1.771

last

mean	0.8569	0.0581	0.3454	0.3156
std. dev.	0.3502	0.2338	0.4755	0.4648

Time taken to build model (full training data) : 170.57 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	2795 ( 10%)	<b>SEQUENTIAL COMPLETER</b>
1	15392 ( 55%)	<b>EARLY DROPOUT</b>
2	6470 ( 23%)	<b>LATE DROPOUT</b>
3	3301 ( 12%)	<b>SAMPLER COMPLETER</b>

Log likelihood: 12.29855

## Principal Component Analyses

The following pages show the output to Principal Component Analyses (PCA) performed on MOOC and PeerWise data post-feature engineering, as performed in this Thesis. A single PCA including all the engineered features for all courses is not feasible, as data across the courses do not have the same number of features, due to differences in learning design, even across different offerings of the same course, as discussed in Section 3.2.2.

The following are results from my preliminary exploration across four of the seventeen courses under consideration in this thesis, namely, the first run of the Understanding Language MOOC (labelled below as UL-1) which had the largest number of enrolments (58,782) and active participants (27,958), the sixth run of Portus (labelled P-6) which had the lowest (7,773 and 5,077, respectively), and both PeerWise courses (12710) and (14715) which mainly differed on the fact that participation was not rewarded by marks (hence, different engagement behaviour was to be expected).

The most relevant features for each course, as reported from PCA are:

**UL-1** num\_comments, LP, num\_steps, step\_completed\_ratio, AR,late\_AR, FR, early\_AR

**P-6** num\_comments, late\_SP, IR, LP, AR, late\_AR, precede, num\_steps

**12710** Comments\_made,First\_comments (FR),b10\_Commentator,Answers\_given, late\_Answer,Initiators\_replies (IR), late\_Ratings, Replies\_made

**14715** Comments\_made,First\_comments,b10\_Commentator,Answers\_given,late\_Answer, Initiators\_replies (IR),late\_Ratings, Replies\_made.

## I.1 PCA for Understanding Language (run 1)

=== Run information ===

```
Evaluator:   weka.attributeSelection.PrincipalComponents -R 0.95 -A 5
Search:      weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 5
Relation:    Experiments/understanding-language1_-weka.filters.unsupervised.attribute.Remove-R1
Instances:   27958
Attributes:  72
             num_steps
             precede
             overlap
             during
             abandoned
             equal
             finish
             meet
             last
             num_comments
             SP
             LP
             FR
             IR
             AR
             pre_abandoned
             pre_begin
             pre_during
             n_pre_start_finish
             pre_last
             pre_meet
             pre_overlap
             pre_precede
             early_abandoned
             early_begin
             early_during
             n_early_finish
             early_last
             early_meet
             early_overlap
             early_precede
             late_abandoned
             late_begin
             late_during
             n_late_finish
             late_last
             late_meet
             late_overlap
             late_precede
             post_abandoned
             post_begin
             post_during
             n_post_end_finish
             post_last
             post_meet
             post_overlap
             post_precede
             pre_AR
             pre_FR
             pre_IR
             pre_LP
             pre_SP
             early_AR
             early_FR
             early_IR
             early_LP
             early_SP
             late_AR
             late_FR
             late_IR
             late_LP
             late_SP
             post_AR
             post_FR
             post_IR
             post_LP
             post_SP
             B1
             B4
```

```

B5
  steps_completed_ratio
  eligible_for_certificate
Evaluation mode:    evaluate on all training data

```

```
=== Attribute Selection on all input data ===
```

```
Search Method:
Attribute ranking.
```

```
Attribute Evaluator (unsupervised):
Principal Components Attribute Transformer
```

```
Correlation matrix (omitted in this view)
```

eigenvalue	proportion	cumulative	
13.68313	0.20122	0.20122	-0.245num_comments-0.227SP-0.224LP-0.218num_steps-0.218steps_completed_ratio...
4.67473	0.06875	0.26997	-0.295AR-0.266late_AR-0.242FR-0.233early_AR-0.223early_FR...
2.92573	0.04303	0.31299	-0.309post_precede-0.265meet-0.261post_last+0.239overlap-0.235post_meet...
2.79687	0.04113	0.35412	-0.327post_LP-0.273post_precede-0.27post_during-0.261post_SP-0.238post_last...
2.43221	0.03577	0.38989	0.239pre_SP+0.236B5+0.236B4-0.228late_FR-0.204FR...
2.3071	0.03393	0.42382	-0.432abandoned-0.311late_abandoned-0.291pre_LP-0.285pre_precede-0.272pre_SP...
2.17033	0.03192	0.45574	0.353abandoned+0.268late_abandoned+0.229early_abandoned-0.227finish-0.225pre_during...
1.99244	0.0293	0.48504	0.613finish+0.437n_pre_start_finish+0.43 n_early_finish+0.143meet+0.134IR...
1.90921	0.02808	0.51311	-0.251IR-0.248early_IR+0.247finish+0.234FR+0.227late_FR...
1.58739	0.02334	0.53646	-0.32early_last-0.296early_precede-0.251early_LP-0.247early_meet+0.236late_precede...
1.47349	0.02167	0.55813	0.302early_last-0.286meet-0.267late_meet+0.26 last+0.24 post_AR...
1.36445	0.02007	0.57819	0.383post_SP+0.343post_FR+0.307post_IR-0.291post_meet+0.224pre_meet...
1.32344	0.01946	0.59765	0.515late_begin+0.51 early_begin-0.233pre_during+0.215last+0.197pre_begin...
1.20007	0.01765	0.6153	0.497pre_last+0.345last+0.231pre_meet+0.23 early_last-0.197post_IR...
1.18415	0.01741	0.63272	-0.353pre_AR-0.344pre_IR-0.255last-0.253early_last-0.244post_IR...
1.10801	0.01629	0.64901	-0.493post_overlap-0.454post_begin+0.334pre_begin+0.261pre_overlap+0.241post_meet...
1.06036	0.01559	0.6646	-0.394pre_begin-0.377post_overlap-0.343pre_overlap+0.261pre_meet-0.217post_begin...
1.04235	0.01533	0.67993	0.355post_begin+0.326post_abandoned+0.298during+0.277pre_during+0.231early_during...
1.00722	0.01481	0.69475	-0.364early_last-0.301post_begin-0.243post_abandoned-0.201post_meet-0.21late_LP...
0.9938	0.01461	0.70936	-0.524n_early_finish+0.515n_pre_start_finish-0.412post_abandoned-0.168post_IR-0.155pre_overlap...
0.98302	0.01446	0.72382	0.579pre_abandoned-0.401post_abandoned-0.252pre_last+0.236n_early_finish-0.232n_pre_start_finish...
0.95813	0.01409	0.73791	-0.547pre_abandoned+0.377early_abandoned-0.318post_abandoned+0.248post_begin-0.212pre_AR...
0.95178	0.014	0.7519	0.49 pre_last+0.404post_begin-0.365post_abandoned+0.278pre_AR-0.232early_last...
0.93264	0.01372	0.76562	-0.324post_begin+0.306pre_begin-0.245pre_abandoned-0.237early_overlap+0.23 during...
0.90612	0.01333	0.77894	0.47 pre_begin+0.401pre_meet+0.326pre_AR+0.288post_begin+0.263post_IR...
0.88997	0.01309	0.79203	0.525post_IR+0.368pre_AR-0.312pre_IR-0.295pre_begin-0.279post_FR...
0.8565	0.0126	0.80463	-0.607early_abandoned+0.601late_abandoned-0.22late_last-0.171post_abandoned-0.159pre_meet...
0.83779	0.01232	0.81695	-0.361pre_meet+0.312post_IR+0.297pre_last-0.266pre_AR+0.257early_during...
0.81059	0.01192	0.82887	-0.443post_AR+0.346pre_IR+0.259early_AR-0.245B5-0.245B4...
0.78592	0.01156	0.84043	0.43 early_meet-0.353late_meet+0.264late_overlap-0.235early_begin-0.221late_abandoned...
0.75996	0.01118	0.8516	-0.43post_AR-0.408pre_IR+0.281early_AR+0.233pre_LP-0.23pre_overlap...
0.72919	0.01072	0.86232	0.356early_AR-0.355late_IR-0.285early_begin+0.272late_begin-0.23early_overlap...
0.71337	0.01049	0.87282	0.499post_FR+0.453post_meet+0.278post_overlap-0.262post_last+0.22 late_last...
0.66347	0.00976	0.88257	0.393late_last+0.317early_meet-0.277post_FR-0.273late_meet-0.271B1...
0.64268	0.00945	0.89202	0.31 early_SP+0.3 post_during+0.299late_begin-0.284B1-0.253early_begin...
0.62076	0.00913	0.90115	-0.49post_SP+0.437post_FR+0.33 post_during+0.284early_AR+0.242early_begin...
0.60856	0.00895	0.9101	-0.416B1-0.373post_during+0.333post_last-0.257late_last-0.243late_IR...
0.59699	0.00878	0.91888	-0.436pre_during+0.411pre_overlap+0.338late_during-0.3post_last-0.266early_overlap...
0.56726	0.00834	0.92722	0.397post_last-0.358post_during+0.345B1+0.245late_last-0.227early_last...
0.55843	0.00821	0.93544	0.353early_during-0.336late_overlap+0.313late_begin-0.287B1+0.274pre_overlap...
0.52257	0.00768	0.94312	-0.409B1-0.317late_SP+0.298pre_LP-0.263pre_FR+0.236pre_precede...
0.51187	0.00753	0.95065	-0.568pre_FR+0.517pre_SP+0.239pre_AR-0.205pre_IR-0.183post_AR...

```
Ranked attributes:
```

```
0.799 1 -0.245num_comments-0.227SP-0.224LP-0.218num_steps-0.218steps_completed_ratio...
```

```
0.73 2 -0.295AR-0.266late_AR-0.242FR-0.233early_AR-0.223early_FR... =>CUM.VAR.=0.29
```

```
0.687 3 -0.309post_precede-0.265meet-0.261post_last+0.239overlap-0.235post_meet...
0.646 4 -0.327post_LP-0.273post_precede-0.27post_during-0.261post_SP-0.238post_last...
0.61 5 0.239pre_SP+0.236B5+0.236B4-0.228late_FR-0.204FR...
```

```
Selected attributes: 1,2,3,4,5 : 5
```

```
Eigenvectors (omitted in this view, available on request)
```



## I.2 PCA on Portus (run 6)

=== Run information ===

```

Evaluator:   weka.attributeSelection.PrincipalComponents -R 0.95 -A 5
Search:     weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 5
Relation:   it/PeerWise/Experiments/portus6_-weka.filters.unsupervised.attribute.Remove-R1
Instances:  969
Attributes: 68
            num_steps
            precede
            overlap
            during
            abandoned
            equal
            finish
            meet
            last
            num_comments
            SP
            LP
            FR
            IR
            AR
            pre_abandoned
            pre_begin
            pre_during
            pre_last
            pre_meet
            pre_overlap
            pre_precede
            early_abandoned
            early_begin
            early_during
            early_last
            early_meet
            early_overlap
            early_precede
            late_abandoned
            late_begin
            late_during
            late_last
            late_meet
            late_overlap
            late_precede
            post_abandoned
            post_begin
            post_during
            post_last
            post_meet
            post_overlap
            post_precede
            pre_AR
            pre_FR
            pre_IR
            pre_LP
            pre_SP
            early_AR
            early_FR
            early_IR
            early_LP
            early_SP
            late_AR
            late_FR
            late_IR
            late_LP
            late_SP
            post_AR
            post_FR
            post_IR
            post_LP
            post_SP
            B1
            B4
            B5
            steps_completed_ratio
            eligible_for_certificate
Evaluation mode:  evaluate on all training data

```

=== Attribute Selection on all input data ===

Search Method:  
Attribute ranking.

Attribute Evaluator (unsupervised):  
Principal Components Attribute Transformer

Correlation matrix (omitted in this view)

eigenvalue	proportion	cumulative	
14.65745	0.22902	0.22902	-0.245num_comments-0.224SP-0.213late_SP-0.212IR-0.205LP...
5.87096	0.09173	0.32076	0.305precede+0.29 num_steps+0.29 steps_completed_ratio+0.27 late_precede+0.237last...
3.92651	0.06135	0.38211	0.282post_AR+0.233post_FR-0.217early_LP+0.215late_FR+0.21 FR...
3.11178	0.04862	0.43073	-0.294post_LP-0.265post_SP-0.235post_during-0.235post_overlap-0.216post_IR...
2.91695	0.04558	0.47631	-0.344overlap-0.319early_overlap+0.301post_precede-0.291late_overlap+0.287post_last...
2.27896	0.03561	0.51192	0.374abandoned+0.33 post_overlap+0.286post_begin+0.281late_abandoned+0.275early_abandoned...
2.12384	0.03319	0.5451	0.34 abandoned-0.313meet+0.288late_abandoned+0.276early_abandoned-0.225late_meet...
2.00884	0.03139	0.57649	0.426meet+0.327early_meet+0.319late_meet+0.262post_SP+0.204post_LP...
1.75152	0.02737	0.60386	0.489early_begin+0.393early_overlap+0.343early_last-0.238pre_meet-0.228pre_during...
1.59109	0.02486	0.62872	0.364post_begin-0.298post_last-0.298pre_last-0.289post_precede+0.285post_overlap...
1.52462	0.02382	0.65254	-0.335post_begin-0.313pre_precede-0.294post_overlap-0.27pre_last-0.255pre_meet...
1.34954	0.02109	0.67363	0.363pre_overlap+0.264IR+0.261late_IR-0.253pre_AR+0.253post_IR...
1.29104	0.02017	0.6938	-0.415pre_meet-0.359pre_last+0.291early_last+0.287early_meet+0.265pre_LP...
1.25016	0.01953	0.71333	-0.33early_meet-0.296during-0.278early_last+0.26 early_begin+0.259pre_overlap...
1.15821	0.0181	0.73143	-0.366pre_IR+0.298B4+0.298B5+0.254pre_AR+0.248pre_overlap...
1.12983	0.01765	0.74908	0.346pre_AR+0.333pre_FR-0.319post_abandoned-0.3post_meet-0.29B4...
1.0311	0.01611	0.76519	0.421post_abandoned-0.398pre_IR+0.297pre_abandoned-0.252early_abandoned+0.213pre_precede...
1.00621	0.01572	0.78092	0.897late_begin-0.176pre_overlap-0.148post_abandoned-0.126pre_IR-0.117pre_during...
0.99743	0.01558	0.7965	-0.35pre_overlap-0.335pre_AR-0.311late_begin+0.301pre_abandoned+0.292pre_LP...
0.97101	0.01517	0.81167	-0.447pre_abandoned+0.383pre_AR-0.336pre_overlap-0.22late_begin-0.207early_LP...
0.90193	0.01409	0.82577	0.501post_abandoned-0.433pre_during+0.206pre_AR-0.191early_during+0.186pre_meet...
0.89206	0.01394	0.8397	0.514pre_abandoned+0.511pre_IR+0.273B4+0.273B5-0.202early_abandoned...
0.82398	0.01287	0.85258	0.368last+0.368pre_during+0.288late_last-0.261pre_meet+0.255early_last...
0.72869	0.01139	0.86396	0.448post_meet-0.322post_abandoned-0.314late_meet-0.306post_during+0.289pre_AR...
0.70437	0.01101	0.87497	-0.539post_SP+0.396post_FR+0.21 early_abandoned-0.208early_FR+0.2 post_LP...
0.65422	0.01022	0.88519	0.412B1+0.323early_last+0.302pre_FR+0.285late_abandoned-0.279early_meet...
0.64051	0.01001	0.8952	0.499pre_last-0.374pre_meet-0.37early_last+0.344early_meet-0.231pre_precede...
0.62286	0.00973	0.90493	0.352late_last-0.308late_abandoned+0.282early_abandoned-0.251late_during+0.245post_during...
0.59955	0.00937	0.9143	0.388post_meet-0.319late_meet-0.269early_during+0.267early_meet-0.256post_last...
0.5569	0.0087	0.923	-0.543late_abandoned+0.436early_abandoned+0.225post_SP-0.222late_last-0.219post_FR...
0.54422	0.0085	0.93151	0.399pre_meet+0.382early_meet-0.359pre_last-0.317post_during+0.246post_last...
0.47559	0.00743	0.93894	0.403B1+0.376pre_during-0.32early_during+0.297post_IR-0.279pre_precede...
0.45908	0.00717	0.94611	0.532pre_precede-0.373pre_SP+0.227late_LP-0.226pre_meet-0.226early_SP...
0.41007	0.00641	0.95252	0.534pre_SP+0.457B1-0.253early_IR+0.219post_IR-0.192pre_during...

Ranked attributes:

0.771 1 -0.245num\_comments-0.224SP-0.213late\_SP-0.212IR-0.205LP...

0.679 2 0.305precede+0.29 num\_steps+0.29 steps\_completed\_ratio+0.27 late\_precede+0.237last... ⇒CUM.VAR.=0.321

0.618 3 0.282post\_AR+0.233post\_FR-0.217early\_LP+0.215late\_FR+0.21 FR...

0.569 4 -0.294post\_LP-0.265post\_SP-0.235post\_during-0.235post\_overlap-0.216post\_IR...

0.524 5 -0.344overlap-0.319early\_overlap+0.301post\_precede-0.291late\_overlap+0.287post\_last...

Selected attributes: 1,2,3,4,5 : 5

Eigenvalues (omitted in this view, available on request)

## I.3 PCA on PeerWise data (course 12710, first cohort of COMP2213)

=== Run information ===

```
Evaluator:   weka.attributeSelection.PrincipalComponents -R 0.95 -A 5
Search:      weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 5
Relation:    12710_full-weka.filters.unsupervised.attribute.Remove-R1
Instances:   135
Attributes:  71
             Group
             User_ID
             Questions_made
             Comments_received
             Starting_questions
             Lone_questions
             Comments_made
             Replies_made
             Initiators_replies
             User
             First_comments
             Followers
             Following
             Ratings_given
             Avg_qual_ratings_given
             Answers_given
             b1-Question_author
             b2-Question_answerer
             b3-Star-crossed
             b4-Comment
             b5-Author-reply
             b6-Follower
             b18-Leader
             b19-Helper
             b23-Verifier
             b7-Good_question_author
             b8-Popular_question_author
             b9-Discussed_question_author
             b10-Commentator
             b11-Critic
             b12-Rater
             b13-Scholar
             b14-Genius
             b15-Einstein
             b16-Insight
             b17-Conversation
             b24-Super_scholar
             b20-I_ll_be_back
             b21-Commitment
             pre_start_Question
             early_Question
             late_Question
             post_end_Question
             pre_start_Answer
             early_Answer
             late_Answer
             post_end_Answer
             pre_start_Comment
             early_Comment
             late_Comment
             post_end_Comment
             pre_start_Ratings
             early_Ratings
             late_Ratings
             post_end_Ratings
             pre_start_Reply
             early_Reply
             late_Reply
             post_end_Reply
             Faculty Code
             Exam_Mark
             Assessment1_Mark
             Assessment2_Mark
             Assessment3_Mark
             Assessment4_Mark
             Assessment4_Late
```

```

Assessment4_Final
Attendance_Mark
Final_Mark
Exam_Mark_nominal
Final_Mark_nominal
Evaluation mode:    evaluate on all training data

```

```
=== Attribute Selection on all input data ===
```

```

Search Method:
Attribute ranking.

```

```

Attribute Evaluator (unsupervised):
Principal Components Attribute Transformer

```

```
Correlation matrix (omitted in this view)
```

```

eigenvalue proportion cumulative
13.80645 0.14688 0.14688 -0.215Comments_made-0.207First_comments-0.204b10-Commentator-0.202Answers_given-0.2b13-Scholar...
6.85922 0.07297 0.21985 0.238late_Answer-0.227Initiators_replies+0.22 late_Ratings-0.216Replies_made+0.214b24-Super_scholar...
4.84389 0.05153 0.27138 -0.33early_Comment-0.309early_Answer-0.279early_Reply-0.276early_Question-0.235b21-Commitment...
4.71649 0.05018 0.32155 -0.288Lone_questions-0.254Assessment4_Mark-0.252Questions_made-0.251Assessment4_Final-0.244b1-Question_author...
4.18085 0.04448 0.36603 0.241pre_start_Comment-0.238Followers-0.223b18-Leader-0.222b8-Popular_question_author-0.197late_Question...
3.75243 0.03992 0.40595 0.277b6-Follower+0.257pre_start_Answer+0.256pre_start_Ratings+0.254Following-0.234Group=group_6...
3.16329 0.03365 0.4396 -0.256late_Question+0.246Assessment1_Mark+0.222Exam_Mark_nominal=first+0.207pre_start_Question-0.16User_ID...
2.80382 0.02983 0.46943 -0.309late_Comment+0.239Comments_received+0.235Starting_questions+0.224pre_start_Ratings-0.215late_Reply...
2.46767 0.02625 0.49568 -0.271Avg_qual_ratings_given-0.217late_Question-0.215Exam_Mark_nominal=first-0.205Exam_Mark+0.201pre_start_Question...
2.4127 0.02567 0.52135 -0.313Assessment4_Late-0.302Group=group_5-0.284Attendance_Mark+0.255Group=group_18+0.242User_ID...
2.23006 0.02372 0.54507 -0.366Group=group_8+0.267Assessment2_Mark-0.254b7-Good_question_author-0.215late_Comment-0.206b4-Comment...
2.09353 0.02227 0.56734 -0.312post_end_Answer-0.277post_end_Ratings-0.271Assessment2_Mark+0.271Group=group_7+0.241post_end_Reply...
1.95224 0.02077 0.58811 -0.417Assessment4_Late-0.29Group=group_5-0.29Group=group_26+0.277Assessment3_Mark-0.25Group=group_18...
1.88236 0.02003 0.60814 -0.303Group=group_4+0.288Group=group_13+0.254Assessment3_Mark+0.238post_end_Reply+0.231Group=group_7...
1.81567 0.01932 0.62745 -0.294Group=group_16+0.249Group=group_5+0.237Group=group_1-0.208post_end_Ratings+0.2 b2-Question_answerer...
1.7121 0.01821 0.64567 -0.404Group=group_25-0.249Assessment1_Mark-0.229Group=group_27+0.219Assessment4_Late+0.208Assessment3_Mark...
1.61529 0.01718 0.66285 0.303Faculty Code=f8+0.292Group=group_6+0.25 pre_start_Ratings+0.224pre_start_Answer+0.219b3-Star-crossed...
1.55895 0.01658 0.67944 0.308Group=group_7+0.307post_end_Reply-0.267Group=group_13+0.244Group=group_18+0.244Exam_Mark_nominal=first...
1.45358 0.01546 0.6949 0.356Group=group_12-0.325Group=group_16-0.257b7-Good_question_author-0.251Group=group_1+0.211User_ID...
1.44544 0.01538 0.71028 0.313Group=group_2-0.299Avg_qual_ratings_given-0.258Group=group_22+0.225Group=group_14+0.213Group=group_24...
1.3273 0.01412 0.7244 0.391Group=group_11+0.264Group=group_18-0.249Group=group_22-0.241Group=group_26-0.214Group=group_8...
1.28914 0.01371 0.73811 0.302Group=group_2+0.262Group=group_11-0.246Group=group_21+0.204Group=group_26+0.2 Avg_qual_ratings_given...
1.27279 0.01354 0.75165 -0.333Group=group_21-0.308Group=group_19-0.264Group=group_25+0.247b23-Verifier-0.222b2-Question_answerer...
1.2325 0.01311 0.76476 0.318Group=group_25-0.289Group=group_3-0.248Group=group_21-0.248Group=group_27+0.234Group=group_22...
1.14344 0.01216 0.77693 0.466Group=group_20-0.32Group=group_14-0.23Group=group_17+0.214Group=group_5-0.213Group=group_26...
1.12391 0.01196 0.78888 0.535Group=group_20+0.25 Group=group_17-0.246Group=group_13+0.244Group=group_21-0.235Group=group_9...
1.10265 0.01173 0.80061 -0.329Group=group_3+0.293Group=group_9-0.275Group=group_24+0.261Group=group_1+0.247Group=group_2...
1.0713 0.0114 0.81201 -0.55Group=group_9-0.42Group=group_21+0.357Group=group_13+0.19 Group=group_4+0.184Group=group_19...
1.02948 0.01095 0.82296 0.504Group=group_3-0.271Group=group_11+0.252Group=group_24-0.237Group=group_27-0.229Group=group_14...
1.02228 0.01088 0.83384 0.433Group=group_4+0.428Group=group_17+0.29 Group=group_13-0.274Group=group_1+0.23 Group=group_9...
0.96065 0.01022 0.84406 0.359Group=group_9+0.311Group=group_19-0.298Group=group_15-0.284Group=group_22-0.203early_Question...
0.92568 0.00985 0.85391 -0.414Group=group_14-0.262User+0.247Group=group_18+0.238Group=group_22-0.203Group=group_26...
0.9066 0.00964 0.86355 0.355Faculty Code=f8-0.317Group=group_1+0.305Group=group_19-0.221late_Reply-0.206early_Ratings...
0.87984 0.00936 0.87291 -0.454Group=group_23+0.241pre_start_Reply-0.223Group=group_14+0.211Group=group_24+0.207Group=group_13...
0.8029 0.00854 0.88145 0.354Group=group_19-0.326Group=group_14+0.314Group=group_17-0.229Group=group_13+0.215Group=group_6...
0.77076 0.0082 0.88965 -0.368Group=group_22+0.293Exam_Mark_nominal=first-0.289post_end_Reply+0.254Group=group_20+0.253Group=group_7...
0.73797 0.00785 0.8975 -0.325Group=group_27+0.228Group=group_2+0.226post_end_Reply+0.212Group=group_6-0.204b23-Verifier...
0.71841 0.00764 0.90514 -0.267Group=group_15+0.223Group=group_24-0.221Group=group_1+0.213pre_start_Ratings+0.207pre_start_Answer...
0.66742 0.0071 0.91225 0.35 b2-Question_answerer+0.233b7-Good_question_author+0.232Group=group_15+0.223early_Reply-0.191Group=group_2...
0.64909 0.00691 0.91915 -0.36early_Reply+0.338Group=group_10-0.275b11-Critic+0.258b2-Question_answerer-0.189Group=group_13...
0.61984 0.00659 0.92574 0.371post_end_Reply-0.296Group=group_7-0.243Group=group_11+0.241Group=group_23-0.22Group=group_4...
0.58772 0.00625 0.932 -0.331b7-Good_question_author+0.288Group=group_24-0.237pre_start_Reply-0.237Group=group_12-0.237Group=group_22...
0.52988 0.00564 0.93763 -0.289b4-Comment+0.27 Faculty Code=f8-0.239b23-Verifier+0.216Group=group_27+0.197Group=group_2...
0.50313 0.00535 0.94299 0.281b23-Verifier+0.281b7-Good_question_author-0.245Group=group_16-0.242Group=group_17+0.229early_Question...
0.48565 0.00517 0.94815 -0.309Group=group_24+0.294b2-Question_answerer-0.267b7-Good_question_author+0.248User_ID+0.245Group=group_10...
0.44412 0.00472 0.95288 0.307Group=group_7+0.278Group=group_10-0.271Group=group_16-0.269Group=group_12-0.207Exam_Mark_nominal=first...

```

```
Ranked attributes:
```

```

0.853 1 -0.215Comments_made-0.207First_comments-0.204b10-Commentator-0.202Answers_given-0.2b13-Scholar...
0.78 2 0.238late_Answer-0.227Initiators_replies+0.22 late_Ratings-0.216Replies_made+0.214b24-Super_scholar...
0.729 3 -0.33early_Comment-0.309early_Answer-0.279early_Reply-0.276early_Question-0.235b21-Commitment...
0.678 4 -0.288Lone_questions-0.254Assessment4_Mark-0.252Questions_made-0.251Assessment4_Final-0.244b1-Question_author...
0.634 5 0.241pre_start_Comment-0.238Followers-0.223b18-Leader-0.222b8-Popular_question_author-0.197late_Question...

```

```
Eigenvectors (omitted in this view, available on request)
```

## I.4 PCA on PeerWise data (course 14715, second cohort of COMP2213)

=== Run information ===

```

Evaluator:   weka.attributeSelection.PrincipalComponents -R 0.95 -A 5
Search:     weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 5
Relation:   14715_full-weka.filters.unsupervised.attribute.Remove-R1
Instances:  106
Attributes: 68
            Group
            User_ID
            Questions_made
            Comments_received
            Starting_questions
            Lone_questions
            Comments_made
            Replies_made
            Initiators_replies
            User
            First_comments
            Followers
            Following
            Ratings_given
            Avg_qual_ratings_given
            Answers_given
            b1-Question_author
            b2-Question_answerer
            b3-Star-crossed
            b4-Comment
            b5-Author-reply
            b6-Follower
            b18-Leader
            b19-Helper
            b23-Verifier
            b7-Good_question_author
            b8-Popular_question_author
            b9-Discussed_question_author
            b10-Commentator
            b11-Critic
            b12-Rater
            b13-Scholar
            b14-Genius
            b15-Einstein
            b16-Insight
            b17-Conversation
            b24-Super_scholar
            b20-I_ll_be_back
            b21-Commitment
            pre_start_Question
            early_Question
            late_Question
            post_end_Question
            pre_start_Answer
            early_Answer
            late_Answer
            post_end_Answer
            pre_start_Comment
            early_Comment
            late_Comment
            post_end_Comment
            pre_start_Ratings
            early_Ratings
            late_Ratings
            post_end_Ratings
            pre_start_Reply
            early_Reply
            late_Reply
            post_end_Reply
            Faculty Code
            Exam_Mark
            Assessment1_Mark
            Assessment2_Mark
            Assessment3_Mark
            Attendance_Mark
            Final_Mark

```

```

Exam_Mark_nominal
Final_Mark_nominal
Evaluation mode:    evaluate on all training data

```

```

=== Attribute Selection on all input data ===

```

```

Search Method:
Attribute ranking.

```

```

Attribute Evaluator (unsupervised):
Principal Components Attribute Transformer

```

```

Correlation matrix (omitted in this view)

```

```

eigenvalue proportion cumulative
19.48617  0.21651  0.21651  -0.203late_Reply-0.201Replies_made-0.192Starting_questions-0.191b18-Leader-0.19b5-Author-reply...
6.56036  0.07289  0.28941  0.251b17-Conversation+0.239Lone_questions+0.227Initiators_replies+0.226late_Question+0.212Questions_made...
5.07735  0.05641  0.34582  0.26  b13-Scholar+0.243b14-Genius+0.238late_Answer+0.238Answers_given-0.221b10-Commentator...
4.03362  0.04482  0.39064  -0.352pre_start_Ratings-0.349pre_start_Answer+0.26  User_ID-0.252Final_Mark+0.226late_Answer...
3.36961  0.03744  0.42808  -0.33Following-0.325pre_start_Question+0.284Assessment1_Mark+0.27  post_end_Ratings+0.268post_end_Answer...
2.93143  0.03257  0.46065  0.353Exam_Mark_nominal=first+0.339Exam_Mark+0.314Final_Mark-0.235Group=group_37-0.214post_end_Ratings...
2.41371  0.02682  0.48747  -0.421early_Answer-0.401early_Ratings-0.379Group=group_20+0.246Assessment1_Mark+0.189b6-Follower...
2.23709  0.02486  0.51233  -0.299Group=group_4+0.279b2-Question_answerer+0.237Group=group_21-0.21post_end_Reply+0.204Answers_given...
1.95461  0.02172  0.53404  -0.336Group=group_18-0.332Faculty Code=f8+0.235post_end_Reply+0.208post_end_Ratings+0.199Group=group_3...
1.83364  0.02037  0.55442  0.31  b19-Helper+0.283Group=group_2-0.261Exam_Mark_nominal=first-0.221Group=group_18+0.221Assessment1_Mark...
1.74418  0.01938  0.5738  0.369Group=group_18+0.344Faculty Code=f8-0.252Group=group_33+0.23  b15-Einstein-0.22late_Ratings...
1.72445  0.01916  0.59296  0.364post_end_Answer+0.343Group=group_33+0.322post_end_Ratings-0.236Avg_qual_ratings_given-0.232post_end_Question...
1.62117  0.01801  0.61097  0.3  User_ID+0.287Group=group_1-0.262Group=group_26+0.232Group=group_22-0.232Group=group_16...
1.53975  0.01711  0.62808  -0.321Group=group_36+0.32  b2-Question_answerer+0.27  b20-I_ll_be_back+0.236Group=group_9+0.219b23-Verifier...
1.49779  0.01664  0.64472  -0.313Group=group_22+0.309Group=group_6-0.278Group=group_10-0.271b6-Follower+0.231b20-I_ll_be_back...
1.40679  0.01563  0.66035  -0.315Group=group_33-0.293Group=group_23-0.275b15-Einstein+0.228post_end_Question+0.21  Group=group_8...
1.37872  0.01532  0.67567  -0.337Group=group_29+0.257Group=group_34+0.239User_ID-0.196Group=group_1+0.192Group=group_24...
1.35093  0.01501  0.69068  -0.436Group=group_29+0.317Group=group_36-0.305b2-Question_answerer+0.276Group=group_19+0.246b23-Verifier...
1.24887  0.01388  0.70456  0.365Group=group_8+0.317Group=group_30-0.279Group=group_9-0.277Group=group_22+0.269Group=group_7...
1.2379  0.01375  0.71831  -0.375Group=group_24+0.359Group=group_8+0.289Group=group_1+0.225Group=group_30+0.216b23-Verifier...
1.17554  0.01306  0.73137  -0.465Group=group_28+0.411Group=group_6-0.262Group=group_9-0.251Group=group_4-0.242Group=group_7...
1.13786  0.01264  0.74402  -0.318Group=group_28-0.318Group=group_27+0.312Group=group_16+0.261Group=group_7-0.257Group=group_2...
1.1274  0.01253  0.75654  -0.453Group=group_9+0.414Group=group_1-0.327Group=group_13-0.292Group=group_8+0.191Group=group_28...
1.08634  0.01207  0.76861  -0.385Group=group_23+0.349Group=group_29+0.315Group=group_34-0.295Group=group_27+0.285Group=group_30...
1.07607  0.01196  0.78057  -0.375Group=group_11-0.313Group=group_36+0.31  Group=group_23+0.29  Group=group_19+0.276Group=group_15...
1.06282  0.01181  0.79238  0.45  Group=group_11-0.402Group=group_2-0.326Group=group_27+0.284Group=group_30+0.243Group=group_35...
1.05697  0.01174  0.80412  -0.484Group=group_13-0.463Group=group_16+0.334Group=group_24-0.261Group=group_19+0.207Group=group_30...
1.03437  0.01149  0.81562  0.549Group=group_24+0.309Group=group_16-0.303Group=group_5+0.279Group=group_19-0.245Group=group_12...
1.03189  0.01147  0.82708  0.622Group=group_26-0.378Group=group_5+0.281Group=group_34+0.25  Group=group_11-0.247Group=group_35...
1.02722  0.01141  0.8385  -0.445Group=group_35+0.384Group=group_19+0.358Group=group_14+0.306Group=group_11-0.297Group=group_26...
1.0256  0.0114  0.84989  -0.69Group=group_5-0.345Group=group_11+0.295Group=group_12+0.279Group=group_15+0.212Group=group_14...
1.01823  0.01131  0.8612  0.638Group=group_14-0.375Group=group_27+0.244Group=group_35+0.232Group=group_13-0.23Group=group_16...
1.01604  0.01129  0.87249  -0.635Group=group_12+0.547Group=group_35-0.237Group=group_16-0.197Group=group_36+0.174Group=group_14...
0.94356  0.01048  0.88298  0.354Group=group_34+0.349Group=group_4-0.262Group=group_15+0.237Group=group_12-0.229Group=group_3...
0.83546  0.00928  0.89226  0.474Group=group_23-0.302Group=group_3-0.296Group=group_21+0.259Group=group_22+0.236post_end_Question...
0.79282  0.00881  0.90107  0.299b19-Helper-0.265Group=group_22-0.258Group=group_34-0.208User-0.207b4-Comment...
0.76586  0.00851  0.90958  0.455Group=group_21-0.34Group=group_37-0.236Group=group_19+0.23  Group=group_15-0.202Group=group_1...
0.73287  0.00814  0.91772  -0.283Group=group_20+0.274Group=group_36+0.237Group=group_19+0.212Group=group_26-0.205Group=group_28...
0.69803  0.00776  0.92548  0.4  b20-I_ll_be_back-0.28Group=group_9+0.238Group=group_2+0.208Group=group_7-0.203pre_start_Answer...
0.64071  0.00712  0.9326  -0.358Group=group_20-0.338User_ID+0.245Group=group_21-0.218b23-Verifier+0.215b20-I_ll_be_back...
0.62196  0.00691  0.93951  -0.277Group=group_37-0.24Group=group_3-0.232b1-Question_author+0.229User+0.216Group=group_18...
0.54111  0.00601  0.94552  0.354b20-I_ll_be_back-0.295Faculty Code=f8+0.265Group=group_37+0.231Group=group_8+0.188User...
0.51804  0.00576  0.95128  0.361Group=group_37-0.336Group=group_20+0.3  Group=group_33+0.248post_end_Question+0.203early_Ratings...

```

```

Ranked attributes:

```

```

0.783  1  -0.203late_Reply-0.201Replies_made-0.192Starting_questions-0.191b18-Leader-0.19b5-Author-reply...
0.711  2  0.251b17-Conversation+0.239Lone_questions+0.227Initiators_replies+0.226late_Question+0.212Questions_made...
0.654  3  0.26  b13-Scholar+0.243b14-Genius+0.238late_Answer+0.238Answers_given-0.221b10-Commentator...
0.609  4  -0.352pre_start_Ratings-0.349pre_start_Answer+0.26  User_ID-0.252Final_Mark+0.226late_Answer...
0.572  5  -0.33Following-0.325pre_start_Question+0.284Assessment1_Mark+0.27  post_end_Ratings+0.268post_end_Answer...

```

```

Selected attributes: 1,2,3,4,5 : 5

```

```

Eigenvectors (omitted in this view, available on request)

```

# Detailed accuracy for classification on clusters found with X-Means

## J.1 Results with $k = 4$

### J.1.1 Portus

=== Run information ===

```
Scheme:          weka.classifiers.sklearn.ScikitLearnClassifier -batch 100 -learner DecisionTreeClassifier
Relation:        DIAL_portus_all_runs-weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.XMeans -I 1 -M 1000
-J 1000 -L 4 -H 4 -B 1.0 -C 0.5 -D "weka.core.EuclideanDistance -R first-last"
-S 10-I6-weka.filters.unsupervised.attribute.Remove-R6
Instances:       16344
Attributes:      6
                 SP
                 LP
                 FR
                 IR
                 AR
                 cluster
Test mode:       10-fold cross-validation
```

=== Classifier model (full training set) ===

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')
```

Time taken to build model: 0.08 seconds

```
=== Stratified cross-validation ===
=== Summary ===
```

Correctly Classified Instances	16279	99.6023 %
Incorrectly Classified Instances	65	0.3977 %
Kappa statistic	0.9781	
Mean absolute error	0.002	
Root mean squared error	0.0446	
Relative absolute error	2.1872 %	
Root relative squared error	20.9238 %	
Total Number of Instances	16344	

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.004	1.000	1.000	1.000	0.996	0.998	0.999	cluster2
	0.978	0.002	0.974	0.978	0.976	0.974	0.988	0.954	cluster1
	0.923	0.002	0.933	0.923	0.928	0.926	0.961	0.863	cluster4
	0.940	0.000	0.969	0.940	0.955	0.954	0.970	0.912	cluster3
Weighted Avg.	0.996	0.004	0.996	0.996	0.996	0.993	0.996	0.993	

```
=== Confusion Matrix ===
```

a	b	c	d	<-- classified as
14735	4	0	0	a = cluster2
7	1070	17	0	b = cluster1
0	25	348	4	c = cluster4
0	0	8	126	d = cluster3

## J.1.2 Understanding Language

```
=== Run information ===
```

```
Scheme:      weka.classifiers.sklearn.ScikitLearnClassifier -batch 100 -learner DecisionTreeClassifier
Relation:    DIAL_understanding-language_all_runs-weka.filters.unsupervised.attribute.Remove-R1
weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.XMeans -I 1 -M 1000
-J 1000 -L 7 -H 7 -B 1.0 -C 0.5 -D "weka.core.EuclideanDistance -R first-last"
-S 10-I6-weka.filters.unsupervised.attribute.Remove-R6
Instances:   121472
Attributes:  6
             SP
             LP
             FR
             IR
             AR
             cluster
Test mode:   10-fold cross-validation
```

```
=== Classifier model (full training set) ===
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')
```



Time taken to build model: 0.27 seconds

=== Stratified cross-validation ===  
 === Summary ===

Correctly Classified Instances	121402	99.9424 %
Incorrectly Classified Instances	70	0.0576 %
Kappa statistic	0.9975	
Mean absolute error	0.0003	
Root mean squared error	0.017	
Relative absolute error	0.2533 %	
Root relative squared error	7.1185 %	
Total Number of Instances	121472	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.999	0.000	0.998	0.999	0.998	0.998	0.999	0.997	cluster4
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	cluster1
	0.992	0.000	0.992	0.992	0.992	0.991	0.996	0.984	cluster3
	0.987	0.000	0.991	0.987	0.989	0.989	0.994	0.979	cluster2
Weighted Avg.	0.999	0.000	0.999	0.999	0.999	0.999	1.000	0.999	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
9279	3	8	0	a = cluster4
2	106270	0	0	b = cluster1
15	0	3893	17	c = cluster3
0	0	25	1960	d = cluster2

### J.1.3 First cohort with PeerWise (12710)

=== Run information ===

```
Scheme:          weka.classifiers.sklearn.ScikitLearnClassifier -batch 100 -learner DecisionTreeClassifier
Relation:        DIAL_12710_4clusters-weka.filters.unsupervised.attribute.Remove-R6
Instances:       135
Attributes:      6
                 SP
                 LP
                 FR
                 IR
                 AR
                 cluster
Test mode:       10-fold cross-validation
```

=== Classifier model (full training set) ===

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')
```

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===  
 === Summary ===

Correctly Classified Instances	130	96.2963 %
Incorrectly Classified Instances	5	3.7037 %
Kappa statistic	0.9379	
Mean absolute error	0.0185	
Root mean squared error	0.1361	
Relative absolute error	6.131 %	
Root relative squared error	35.144 %	
Total Number of Instances	135	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.986	0.015	0.986	0.986	0.986	0.970	0.985	0.979	cluster1
	0.980	0.023	0.960	0.980	0.970	0.952	0.978	0.948	cluster3
	0.889	0.008	0.889	0.889	0.889	0.881	0.940	0.798	cluster4
	0.750	0.008	0.857	0.750	0.800	0.790	0.871	0.658	cluster2
Weighted Avg.	0.963	0.017	0.962	0.963	0.962	0.947	0.973	0.936	

=== Confusion Matrix ===

```

a  b  c  d  <-- classified as
68  1  0  0 | a = cluster1
 1 48  0  0 | b = cluster3
 0  0  8  1 | c = cluster4
 0  1  1  6 | d = cluster2

```

## J.1.4 Second cohort with PeerWise (14715)

=== Run information ===

```

Scheme:      weka.classifiers.sklearn.ScikitLearnClassifier -batch 100 -learner DecisionTreeClassifier
Relation:    DIAL_14715_4clusters-weka.filters.unsupervised.attribute.Remove-R6
Instances:   106
Attributes:  6
             SP
             LP
             FR
             IR
             AR
             cluster
Test mode:   10-fold cross-validation

```

=== Classifier model (full training set) ===

```

DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')

```

Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===  
 === Summary ===

Correctly Classified Instances	103	97.1698 %
Incorrectly Classified Instances	3	2.8302 %
Kappa statistic	0.9248	
Mean absolute error	0.0142	
Root mean squared error	0.119	
Relative absolute error	7.3351 %	
Root relative squared error	39.0171 %	
Total Number of Instances	106	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.988	0.000	1.000	0.988	0.994	0.973	0.994	0.997	cluster4
	1.000	0.010	0.909	1.000	0.952	0.948	0.995	0.909	cluster2
	0.875	0.010	0.875	0.875	0.875	0.865	0.932	0.775	cluster3
	0.800	0.010	0.800	0.800	0.800	0.790	0.895	0.649	cluster1
Weighted Avg.	0.972	0.002	0.973	0.972	0.972	0.954	0.985	0.956	

=== Confusion Matrix ===

```

a  b  c  d  <-- classified as
82  1  0  0 | a = cluster4
 0 10  0  0 | b = cluster2
 0  0  7  1 | c = cluster3
 0  0  1  4 | d = cluster1

```

## J.2 Results with $k = 7$

### J.2.1 Portus

=== Run information ===

```

Scheme:      weka.classifiers.sklearn.ScikitLearnClassifier -batch 100 -learner DecisionTreeClassifier
Relation:    DIAL_portus_all_runs-weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.XMeans -I 1 -M 1000
-J 1000 -L 7 -H 7 -B 1.0 -C 0.5 -D "weka.core.EuclideanDistance -R first-last"
-S 10-I6-weka.filters.unsupervised.attribute.Remove-R6
Instances:   16344
Attributes:  6
             SP
             LP
             FR
             IR
             AR
             cluster
Test mode:   10-fold cross-validation

```

=== Classifier model (full training set) ===

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
    max_features=None, max_leaf_nodes=None,
    min_impurity_decrease=0.0, min_impurity_split=None,
    min_samples_leaf=1, min_samples_split=2,
    min_weight_fraction_leaf=0.0, presort=False, random_state=None,
    splitter='best')
```

Time taken to build model: 0.17 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	16254	99.4493 %
Incorrectly Classified Instances	90	0.5507 %
Kappa statistic	0.9719	
Mean absolute error	0.0016	
Root mean squared error	0.0397	
Relative absolute error	2.795 %	
Root relative squared error	23.6619 %	
Total Number of Instances	16344	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.980	0.002	0.977	0.980	0.979	0.977	0.989	0.959	cluster1
	0.860	0.001	0.854	0.860	0.857	0.856	0.929	0.736	cluster2
	0.688	0.000	0.815	0.688	0.746	0.748	0.844	0.561	cluster3
	0.939	0.000	0.930	0.939	0.934	0.934	0.969	0.874	cluster4
	0.955	0.001	0.949	0.955	0.952	0.951	0.977	0.908	cluster5
	1.000	0.008	0.999	1.000	0.999	0.993	0.996	0.999	cluster6
	0.640	0.000	1.000	0.640	0.780	0.800	0.820	0.641	cluster7
Weighted Avg.	0.994	0.007	0.994	0.994	0.994	0.989	0.994	0.990	

=== Confusion Matrix ===

	a	b	c	d	e	f	g	<-- classified as
1086	6	0	0	3	13	0	0	a = cluster1
8	129	3	1	9	0	0	0	b = cluster2
0	5	22	5	0	0	0	0	c = cluster3
0	1	2	107	4	0	0	0	d = cluster4
5	7	0	2	299	0	0	0	e = cluster5
7	0	0	0	0	14595	0	0	f = cluster6
5	3	0	0	0	1	16	0	g = cluster7

## J.2.2 Understanding Language

=== Run information ===

```
Scheme:      weka.classifiers.sklearn.ScikitLearnClassifier -batch 100 -learner DecisionTreeClassifier
Relation:    DIAL_understanding-language_all_runs_7clusters-weka.filters.unsupervised.attribute.Remove-R6
Instances:   121472
Attributes:  6
             SP
             LP
```

```

FR
IR
AR
cluster
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')

Time taken to build model: 1.56 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      121367           99.9136 %
Incorrectly Classified Instances    105              0.0864 %
Kappa statistic                    0.9978
Mean absolute error                 0.0002
Root mean squared error            0.0157
Relative absolute error            0.222 %
Root relative squared error        6.6638 %
Total Number of Instances         121472

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.999   0.000   0.998     0.999   0.998     0.998   0.999   0.997   cluster7
          1.000   0.000   1.000     1.000   1.000     1.000   1.000   1.000   cluster6
          0.998   0.000   0.997     0.998   0.997     0.997   0.999   0.995   cluster4
          1.000   0.000   1.000     1.000   1.000     1.000   1.000   1.000   cluster5
          0.991   0.000   0.994     0.991   0.993     0.992   0.996   0.985   cluster3
          0.988   0.000   0.980     0.988   0.984     0.984   0.994   0.969   cluster2
          0.939   0.000   0.956     0.939   0.947     0.947   0.969   0.898   cluster1
Weighted Avg.  0.999   0.000   0.999     0.999   0.999     0.999   1.000   0.998

=== Confusion Matrix ===

   a    b    c    d    e    f    g  <-- classified as
7258   1    6    1    0    0    2 |  a = cluster7
 3 93848   0    0    0    0    0 |  b = cluster6
 6    0 4832   0    2    0    3 |  c = cluster4
 0    0  0 10827   0    0    0 |  d = cluster5
 0    0  7    0 2656   3   13 |  e = cluster3
 0    0  0    0  4 1297  12 |  f = cluster2
 5    0  3    0 11  23  649 |  g = cluster1

```

## J.2.3 First cohort with PeerWise (12710)

```
=== Run information ===
```

```

Scheme:      weka.classifiers.sklearn.ScikitLearnClassifier -batch 100 -learner DecisionTreeClassifier
Relation:    DIAL_12710_7clusters-weka.filters.unsupervised.attribute.Remove-R6
Instances:   135
Attributes:  6
            SP
            LP
            FR
            IR
            AR
            cluster
Test mode:   10-fold cross-validation

```

```
=== Classifier model (full training set) ===
```

```

DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')

```

```
Time taken to build model: 0.04 seconds
```

```
=== Stratified cross-validation ===
```

```
=== Summary ===
```

Correctly Classified Instances	126	93.3333 %
Incorrectly Classified Instances	9	6.6667 %
Kappa statistic	0.9202	
Mean absolute error	0.019	
Root mean squared error	0.138	
Relative absolute error	7.9506 %	
Root relative squared error	39.8869 %	
Total Number of Instances	135	

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.931	0.019	0.931	0.931	0.931	0.912	0.956	0.882	cluster3
	1.000	0.009	0.955	1.000	0.977	0.973	0.996	0.955	cluster2
	0.955	0.027	0.875	0.955	0.913	0.896	0.964	0.843	cluster1
	0.895	0.009	0.944	0.895	0.919	0.906	0.943	0.860	cluster4
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	cluster7
	0.429	0.016	0.600	0.429	0.500	0.485	0.706	0.287	cluster6
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	cluster5
Weighted Avg.	0.933	0.012	0.929	0.933	0.930	0.920	0.961	0.885	

```
=== Confusion Matrix ===
```

```

 a b c d e f g  <-- classified as
27 1 1 0 0 0 0 | a = cluster3
 0 21 0 0 0 0 0 | b = cluster2
 0 0 21 0 0 1 0 | c = cluster1
 0 0 1 17 0 1 0 | d = cluster4
 0 0 0 0 26 0 0 | e = cluster7
 2 0 1 1 0 3 0 | f = cluster6
 0 0 0 0 0 0 11 | g = cluster5

```

## J.2.4 Second cohort with PeerWise (14715)

=== Run information ===

```

Scheme:      weka.classifiers.sklearn.ScikitLearnClassifier -batch 100 -learner DecisionTreeClassifier
Relation:    DIAL_14715_7clusters-weka.filters.unsupervised.attribute.Remove-R6
Instances:   106
Attributes:  6
             SP
             LP
             FR
             IR
             AR
             cluster
Test mode:   10-fold cross-validation

```

=== Classifier model (full training set) ===

```

DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')

```

Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	98	92.4528 %
Incorrectly Classified Instances	8	7.5472 %
Kappa statistic	0.8306	
Mean absolute error	0.0216	
Root mean squared error	0.1468	
Relative absolute error	15.4942 %	
Root relative squared error	57.0881 %	
Total Number of Instances	106	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.069	0.975	1.000	0.987	0.953	0.966	0.975	cluster7
	0.667	0.019	0.500	0.667	0.571	0.563	0.824	0.343	cluster4
	0.875	0.020	0.778	0.875	0.824	0.810	0.927	0.690	cluster3
	0.400	0.010	0.667	0.400	0.500	0.499	0.695	0.295	cluster1
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	cluster2
	0.250	0.010	0.500	0.250	0.333	0.336	0.620	0.153	cluster6
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	cluster5
Weighted Avg.	0.925	0.053	0.916	0.925	0.916	0.890	0.936	0.874	

=== Confusion Matrix ===

```

  a  b  c  d  e  f  g  <-- classified as
77  0  0  0  0  0  0  | a = cluster7
  0  2  0  0  0  1  0  | b = cluster4
  0  0  7  1  0  0  0  | c = cluster3
  0  1  2  2  0  0  0  | d = cluster1

```

```
0 0 0 0 5 0 0 | e = cluster2
2 1 0 0 0 1 0 | f = cluster6
0 0 0 0 0 0 4 | g = cluster5
```



## Clustering FutureLearn MOOCs with X-Means and $k=4$

I performed these explorations as part of my interest in validating the model presented in Chapter 4, which includes temporal interval features. As explained in Section 5.6, these results were excluded from the main analysis as I made the decision to focus on results for seven clusters (i.e.  $k=7$  for the X-Means algorithm, achieved by forcing both its L and H parameters to this value). Seven clusters are also closer to Chua's taxonomy in terms of number of distinctively identifiable groups.

Figures K.1 and K.2 show, for each of the two MOOCs under consideration, the distribution of learners across these four clusters. As mentioned in Chapter 5, when letting X-Means choose the best number of clusters for the data between the values of two (parameter L) and ten (parameter H), it typically would return seven clusters on this data. However, these results were not always consistent and on occasion it would return four clusters instead (particularly when the clusterer is used as an instance filter). The reason behind it might be that the difference in performance for X-Means is almost negligible across the different values of  $k$ , as shown in in Figures 5.8 and 5.7.

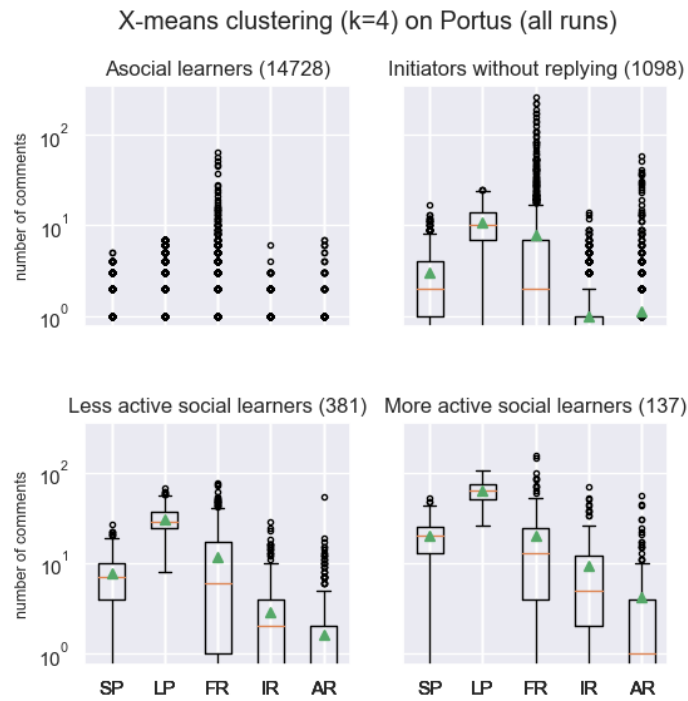


FIGURE K.1: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the Portus MOOC (all runs), with  $k=4$

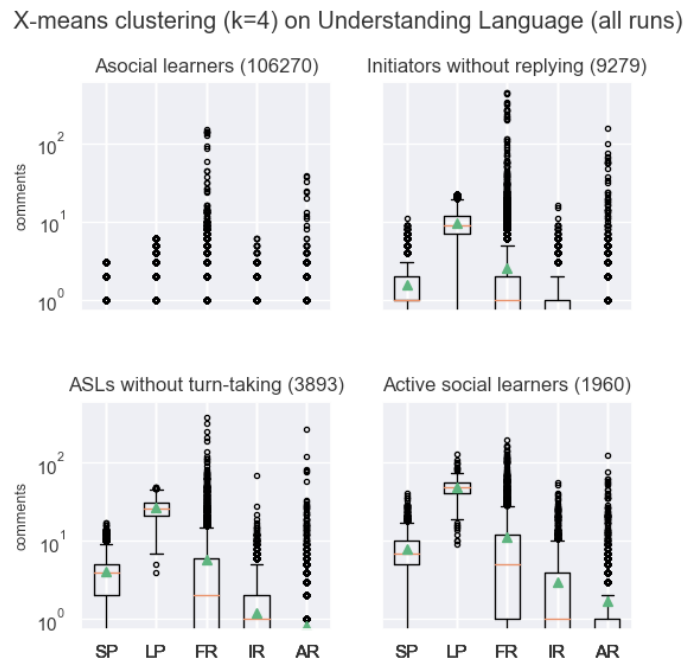


FIGURE K.2: Box-and-whisker plots for the clusters found by the X-Means clustering algorithm on the Understanding Language MOOC (all runs), with  $k=4$

## Clustering individual runs of FutureLearn MOOCs with X-Means and $k=7$

As explained in Section 5.6, there was some variation on the semantics for the seven clusters found for each of the runs under consideration (i.e.  $k=7$  for the X-Means algorithm, achieved by forcing both its L and H parameters to this value). The box-and-whiskers plots for each are shown here, as the interpretation of these was used to generate Tables 5.12 and 5.13 for Portus and Tables 5.14 and 5.15.

The colours for the clusters in the captions for Figures L.2 to L.7 (the boxplots for Portus) and Figures L.9 to L.18 (the boxplots for Understanding Language) facilitated the semantic coding, matching those in Tables 5.14 and 5.15 for Understanding Language.

Also Figures L.1 and L.8 are shown, presenting the confusion matrices associated to the classification of the clusters in each of the runs of each MOOCs. Note that both box-and-whiskers plots and confusion matrices should be interpreted as per the explanations given in Section 5.7.

## L.1 Portus

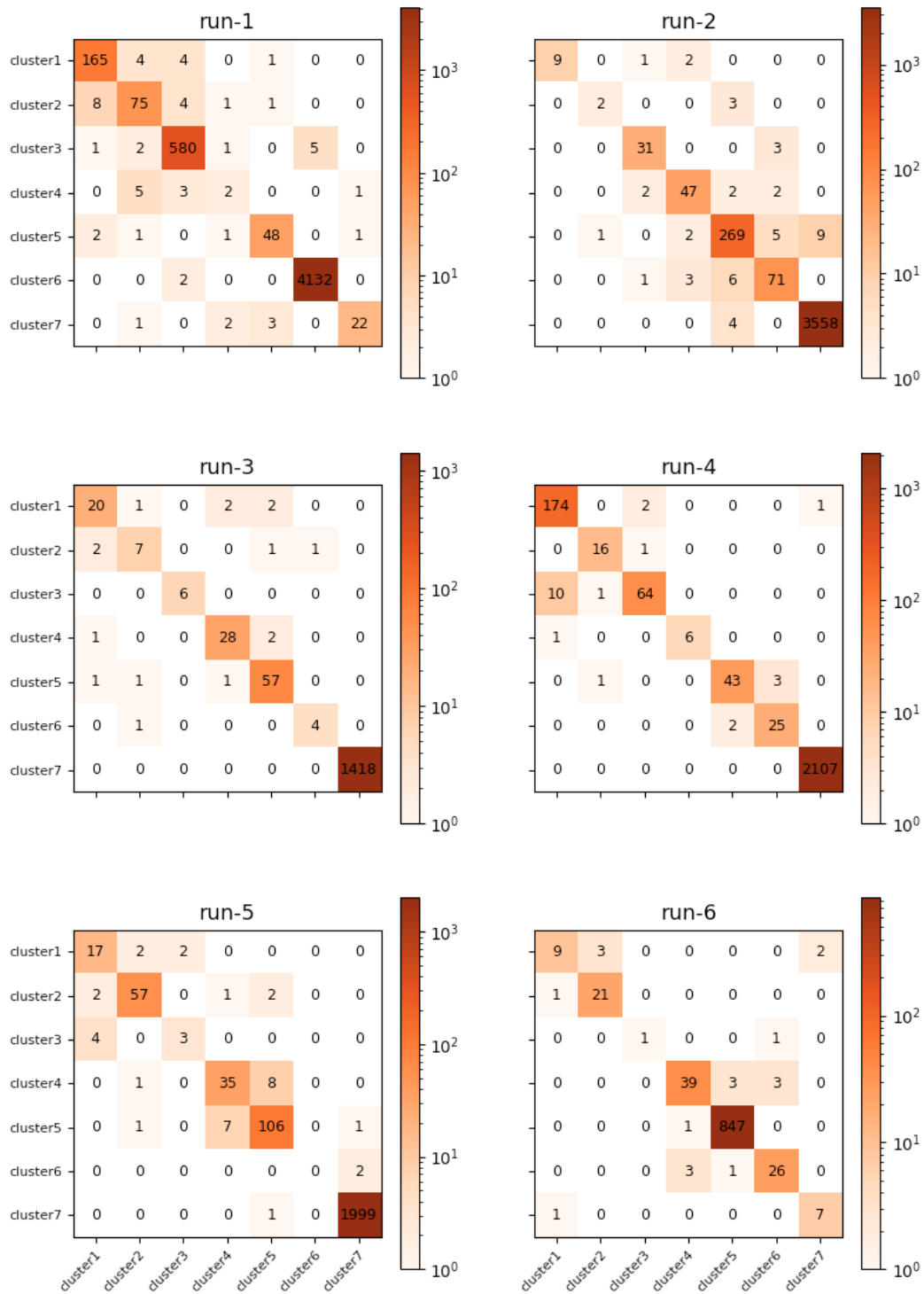


FIGURE L.1: Confusion matrix plots for the clusters found by the X-Means clustering algorithm on the Portus MOOC (for each run), with  $k = 7$ .

## X-means clustering on portus (run 1)

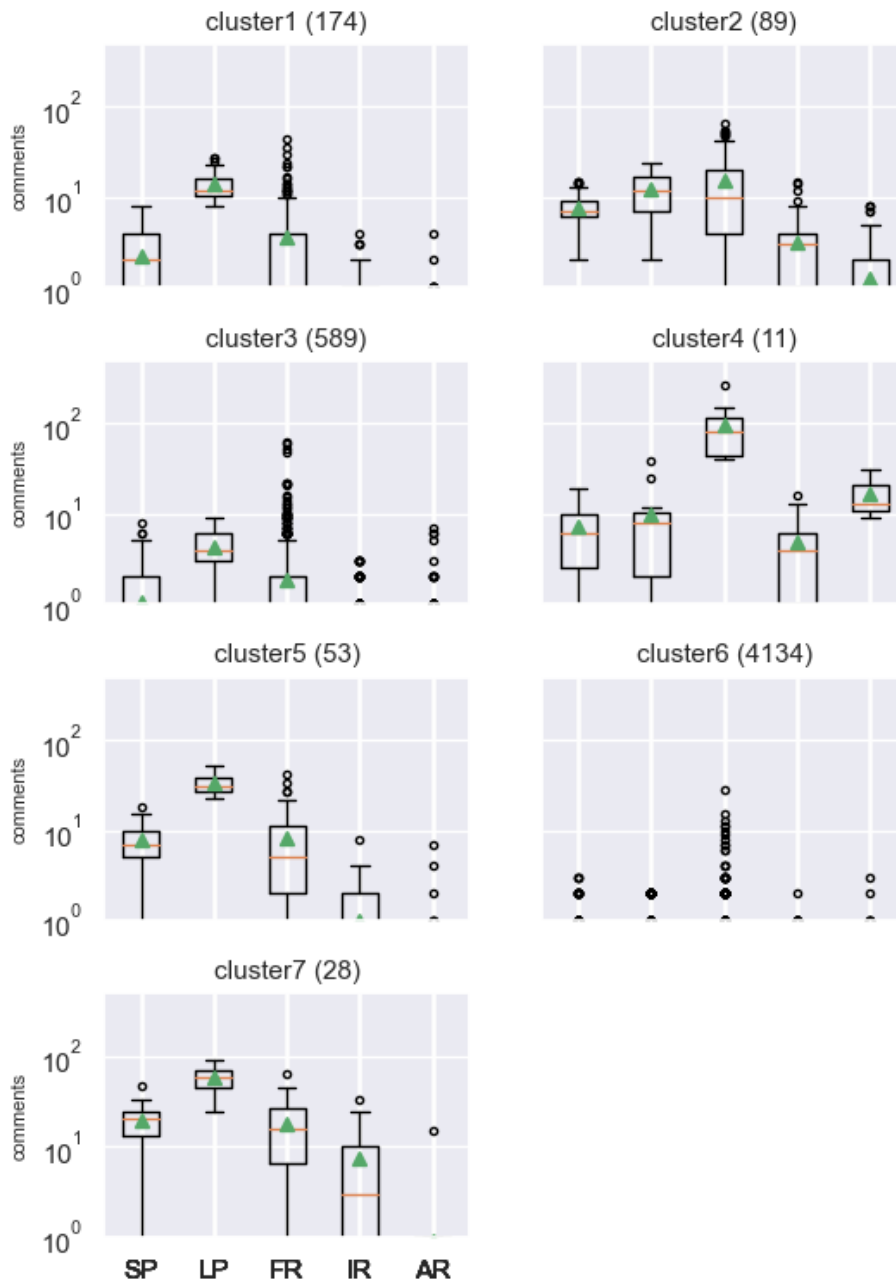


FIGURE L.2: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Portus MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features, are as follows: **cluster1** : 7-ASL without turn-taking (only posting); **cluster2** : 8-active social learners; **cluster3** : 6-reluctant ASL; **cluster4** : 8a-more active social learners; **cluster5** : 7-ASL without turn-taking; **cluster6** : 1-asocial learners; **cluster7** : 4-initiators who respond.

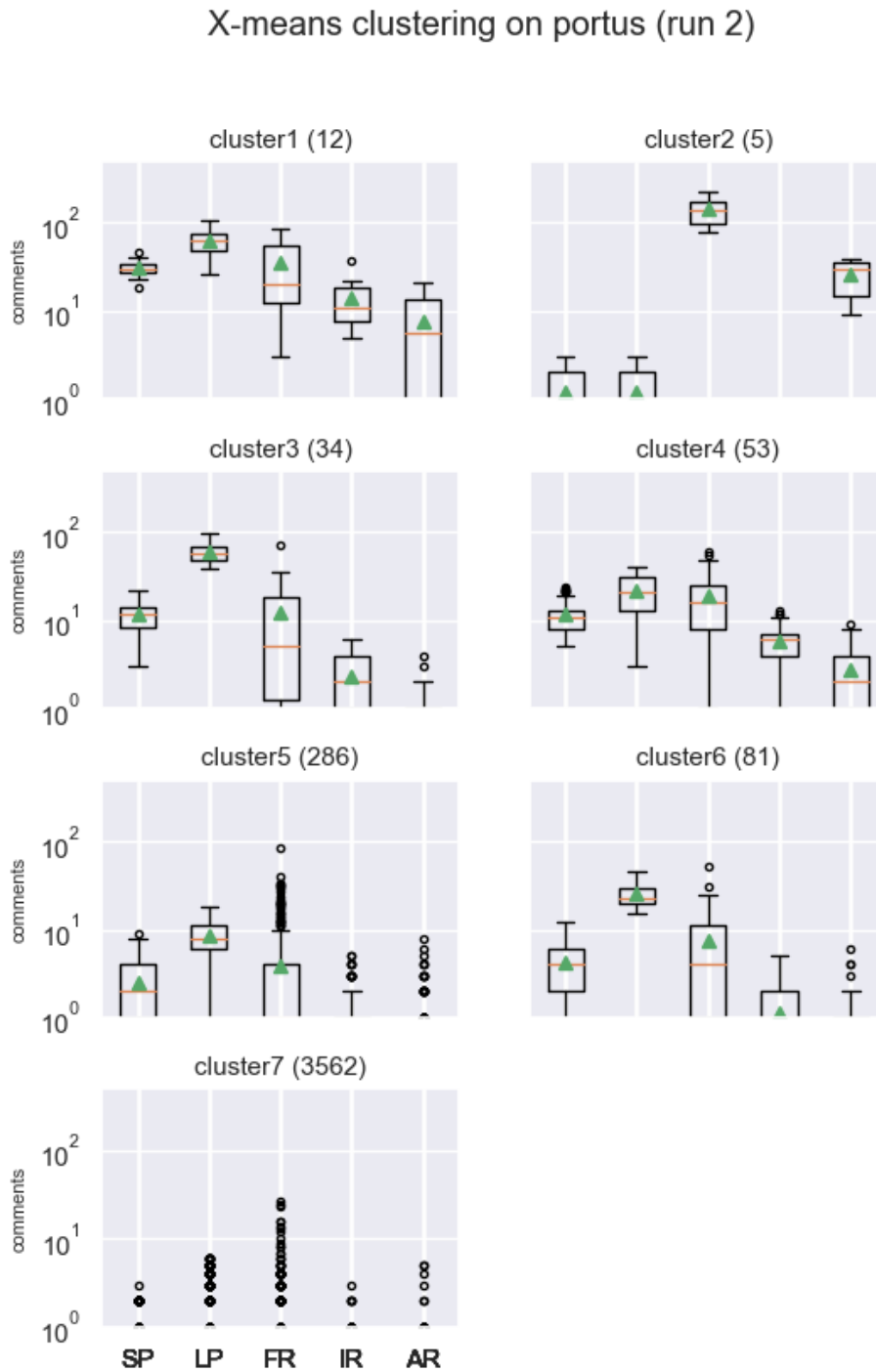


FIGURE L.3: Distribution of dialogic features in clusters found by the X-Means algorithm on the second run of the Portus MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features, are as follows: **cluster1** : 8a-more active social learners; **cluster2** : 5-repliers; **cluster3** : 8b-ASL who do not give additional replies; **cluster4** : 8-active social learners; **cluster5** : 2-loners; **cluster6** : 7-ASL without turn-taking; **cluster7** : 1-asocial learners.

## X-means clustering on portus (run 3)

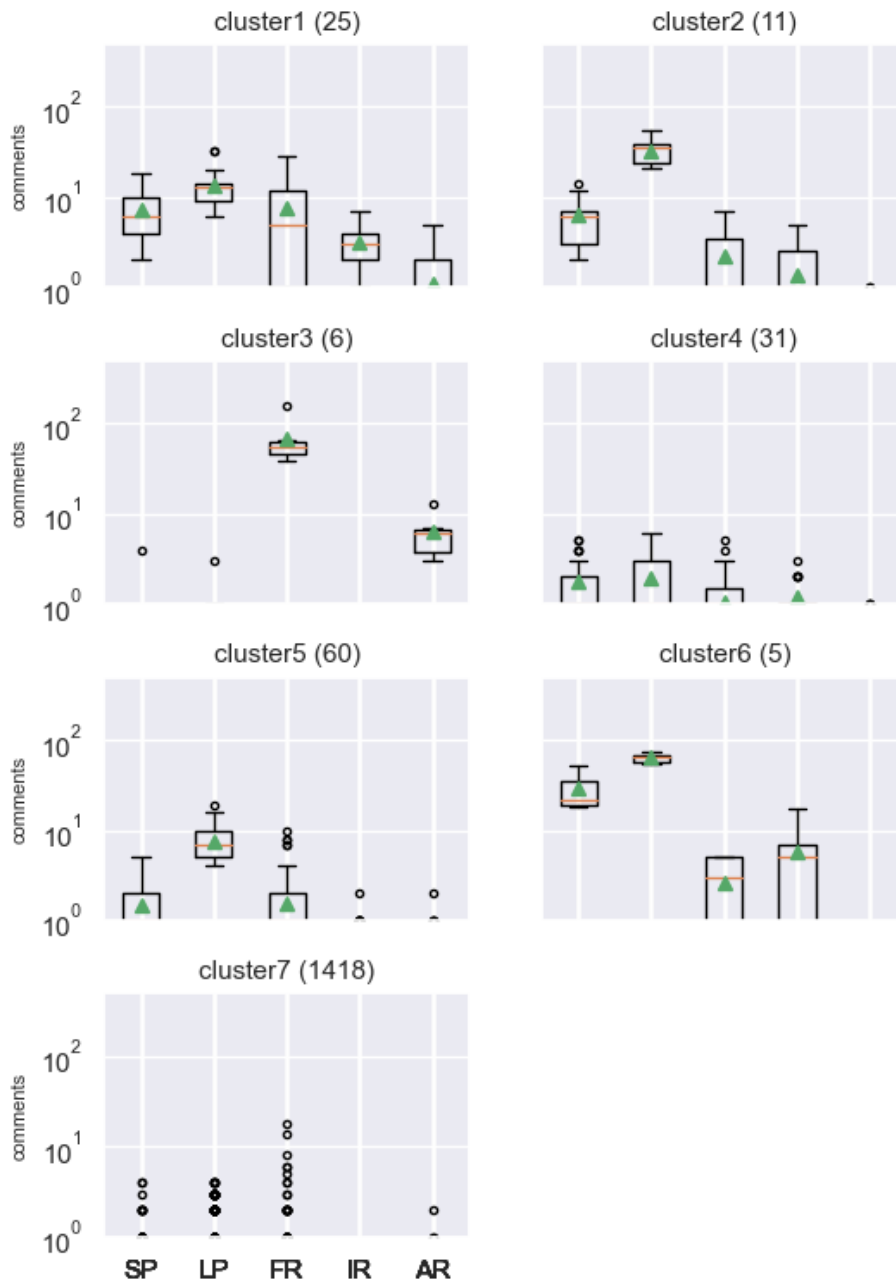


FIGURE L.4: Distribution of dialogic features in clusters found by the X-Means algorithm on the third run of the Portus MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1** : 8-active social learners; **cluster2** : 7-ASL without turn-taking; **cluster3** : 5-replier; **cluster4** : 8-active social learners; **cluster5** : 2-loners; **cluster6** : 8a-more active social learners **cluster7** : 1-asocial learners.

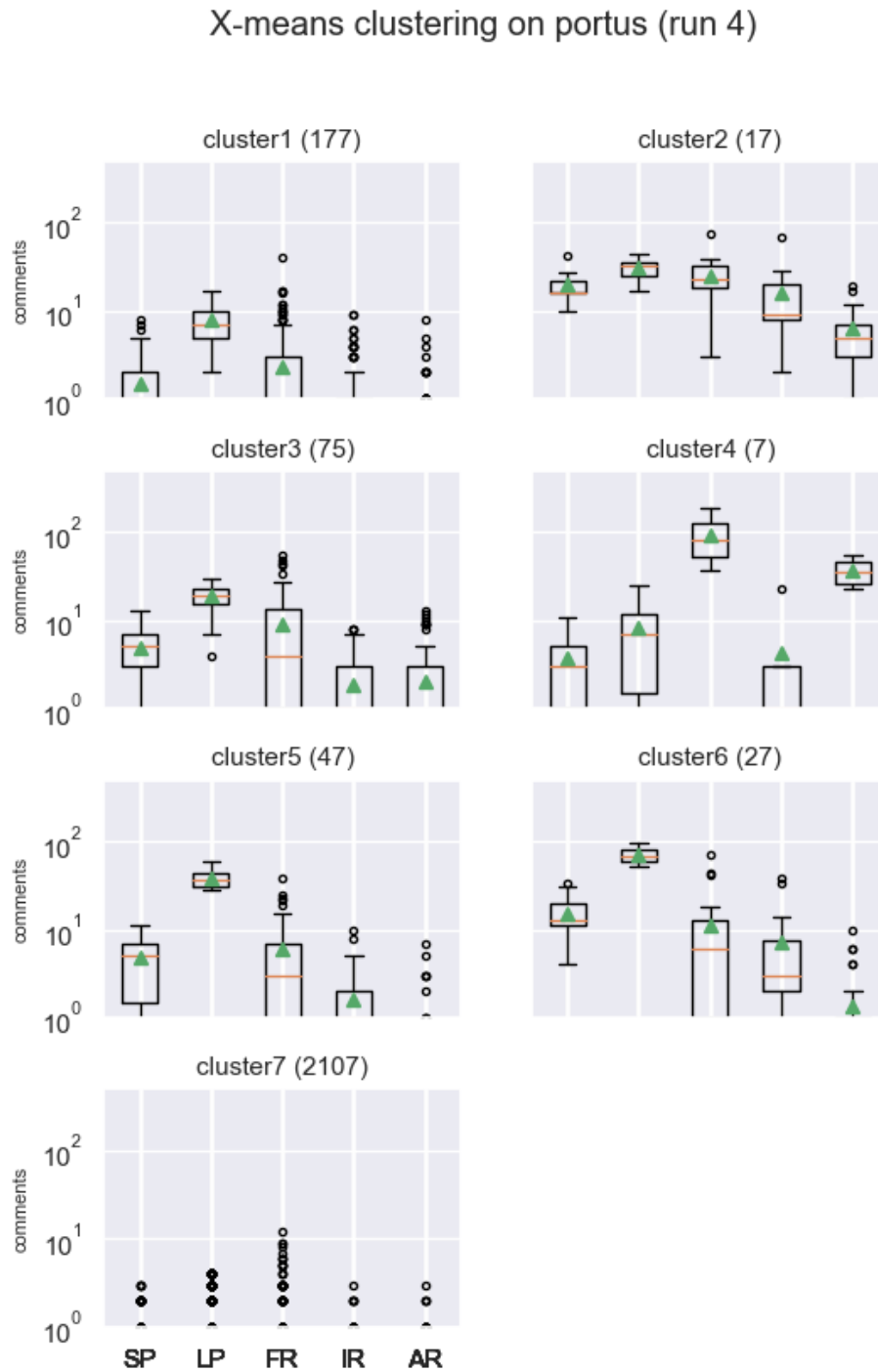


FIGURE L.5: Distribution of dialogic features in clusters found by the X-Means algorithm on the fourth run of the Portus MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1** : 2-loners; **cluster2** : 8aa-even more active social learners; **cluster3** : 8-active social learners; **cluster4** : 8a-more active social learners; **cluster5** : 7-ASL without turn-taking; **cluster6** : 8b-ASL who do not give additional replies; **cluster7** : 1-asocial learners.



## X-means clustering on portus (run 5)

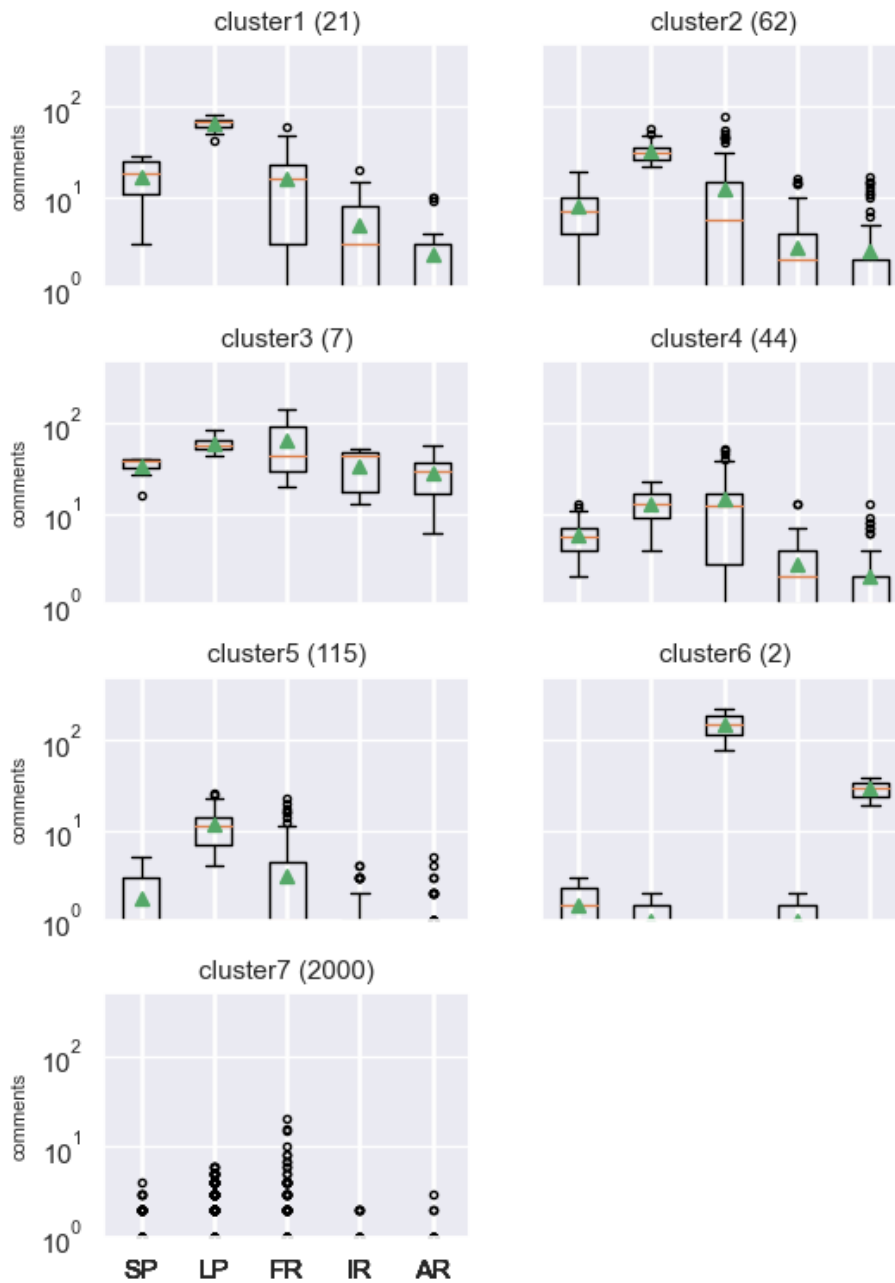


FIGURE L.6: Distribution of dialogic features in clusters found by the X-Means algorithm on the fifth run of the Portus MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1** : 8bb-more active SL who do not give additional replies; **cluster2** : 8b-ASL who do not give additional replies; **cluster3** : 8a-more active social learners; **cluster4** : 8-active social learners; **cluster5** : 6-reluctant ASL; **cluster6** : 5-replier; **cluster7** : 1-asocial learners.

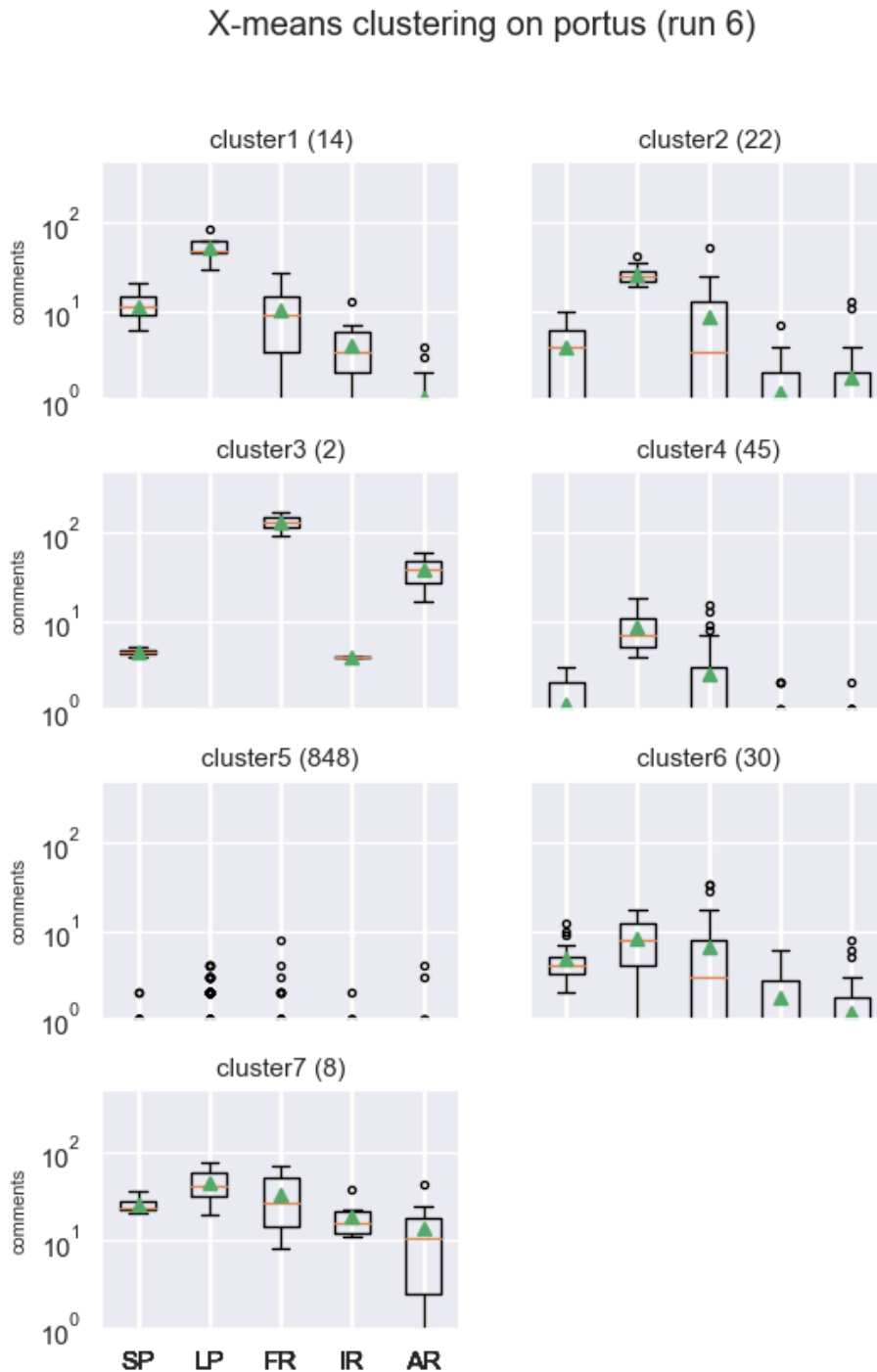


FIGURE L.7: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Portus MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows **cluster1** : 8bb-more active SL who do not give additional replies; **cluster2** : 8b-ASL who do not give additional replies; **cluster3** : 8a-more active social learners; **cluster4** : 2-loners; **cluster5** :1-asocial learners; **cluster6** : 7-ASL without turn-taking; **cluster7** : 8-active social learners.

## L.2 Understanding Language

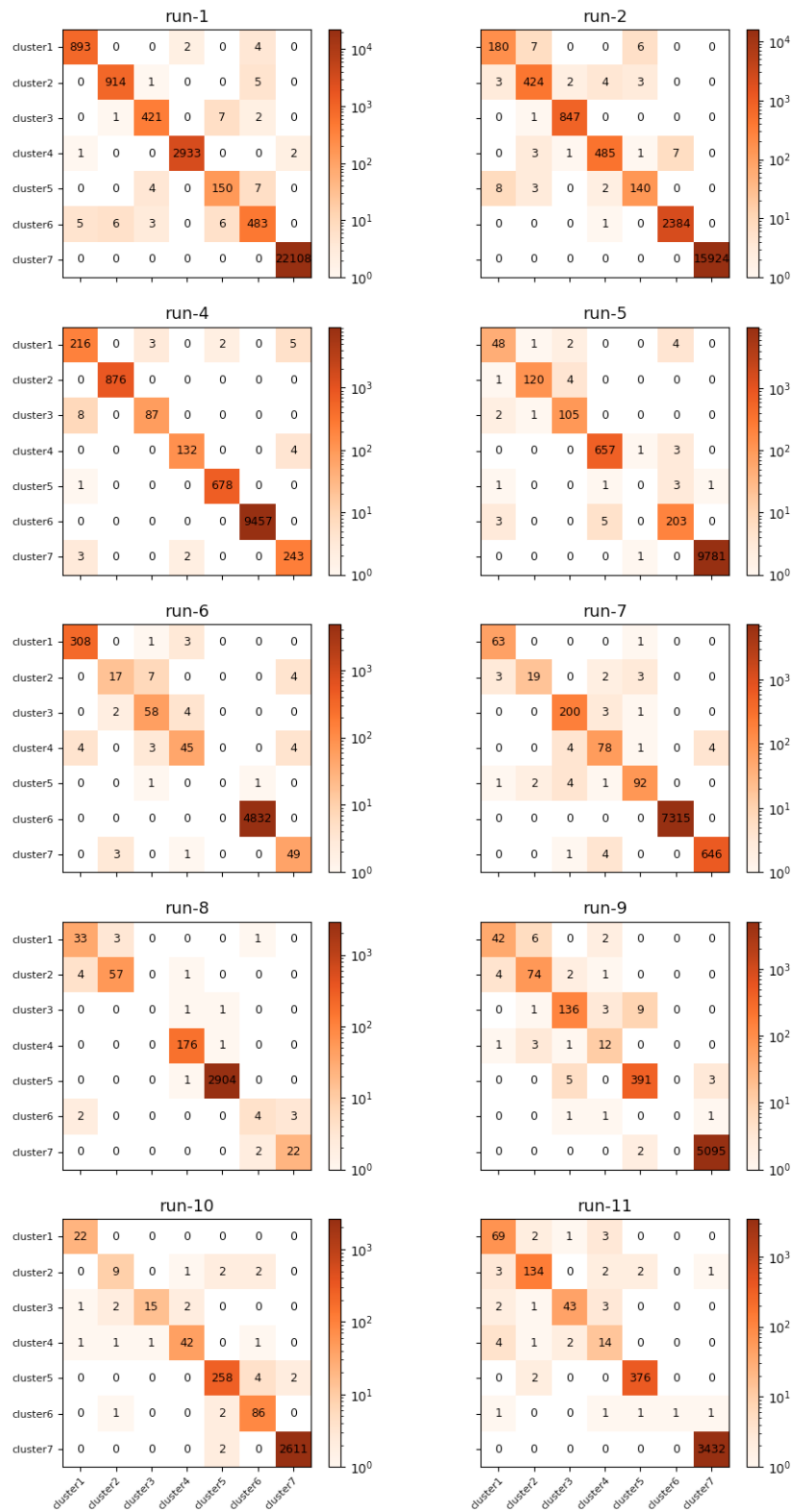


FIGURE L.8: Confusion matrix plots for the clusters found by the X-Means clustering algorithm on the Understanding Language MOOC (for each run), with  $k = 7$ .

## X-means clustering on understanding-language (run 1)

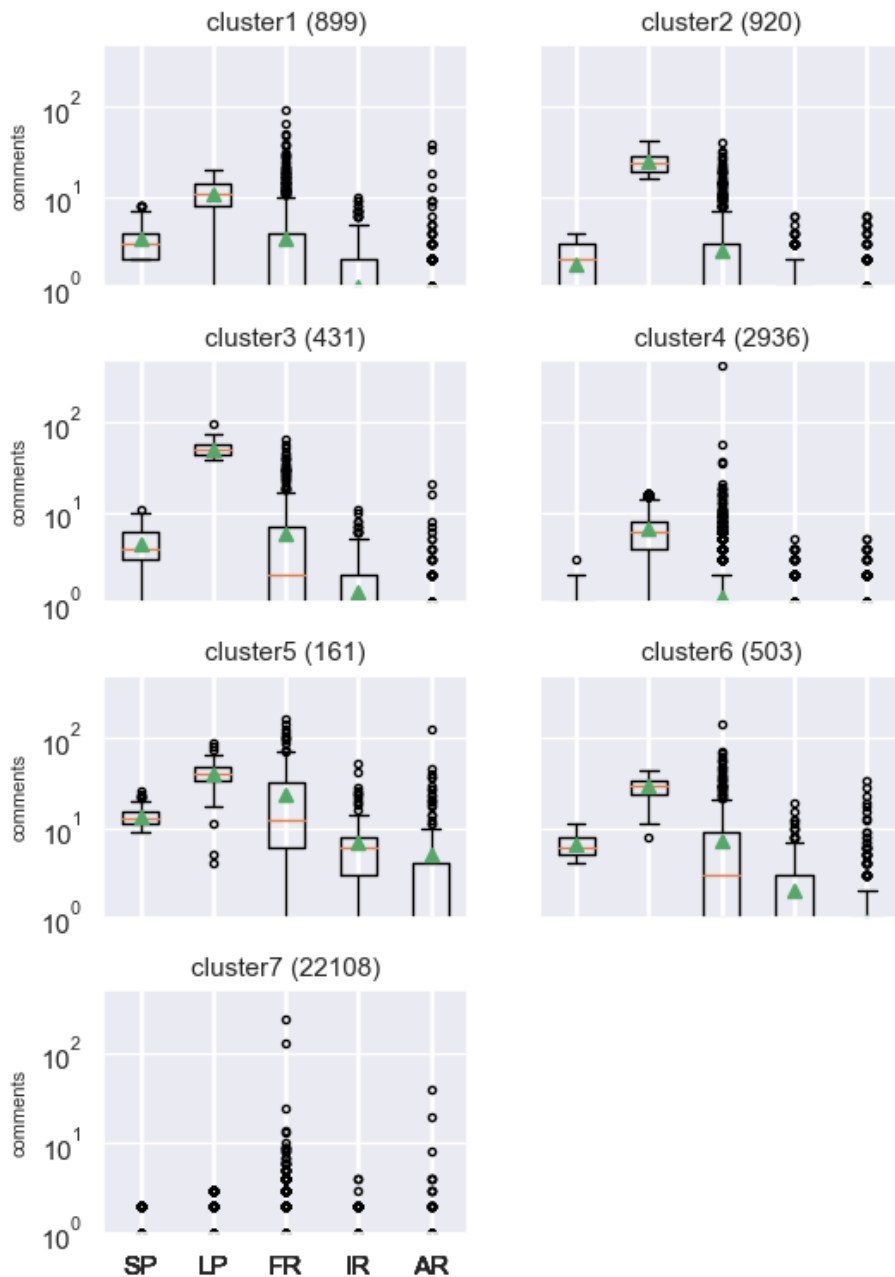


FIGURE L.9: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Understanding Language MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1** : 3a-more active initiators without replying; **cluster2** : 3-initiators without replying; **cluster3** : 4-initiators who respond; **cluster4** : 2-loners; **cluster5** : 8-active social learners; **cluster6** : 7-ASL without turn-takings; **cluster7** : 1-asocial learners.

## X-means clustering on understanding-language (run 2)

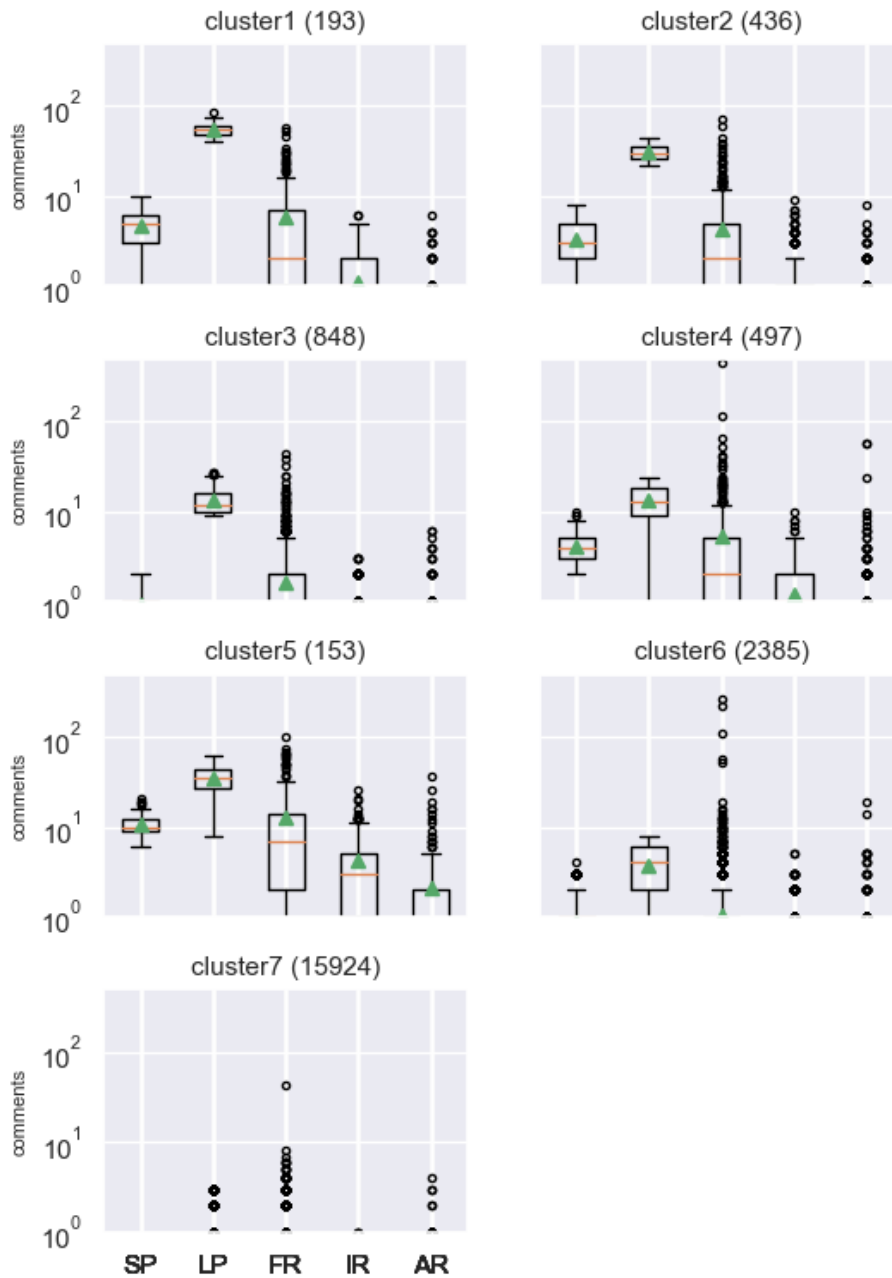


FIGURE L.10: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Understanding Language MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows: **cluster1** : 4-initiators who respond; **cluster2** : 3-initiators without replying; **cluster3** : 2a-more active loners; **cluster4** : 7-ASL without turn-taking; **cluster5** : 8-active social learners; **cluster6** : 2-loners; **cluster7** : 1-asocial learners.

## X-means clustering on understanding-language (run 4)

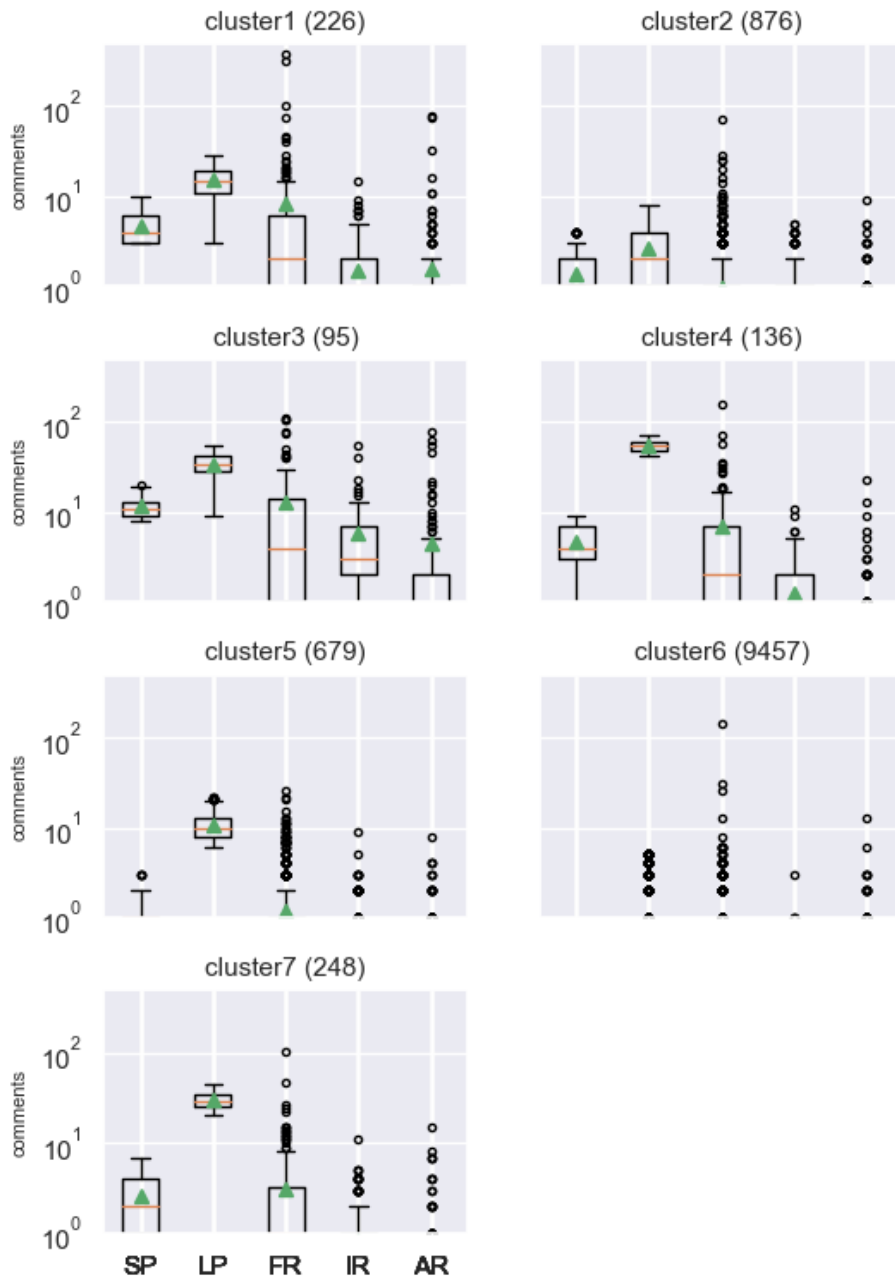


FIGURE L.11: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Understanding Language MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features, are as follows: **cluster1** : 7-ASL without turn-taking; **cluster2** : 3-initiators without replying; **cluster3** : 8-active social learners; **cluster4** : 7a-more active SL without turn-taking; **cluster5** : 2-loners; **cluster6** : 1-asocial learners; **cluster7** : 3a-more active initiators without replying.

## X-means clustering on understanding-language (run 5)

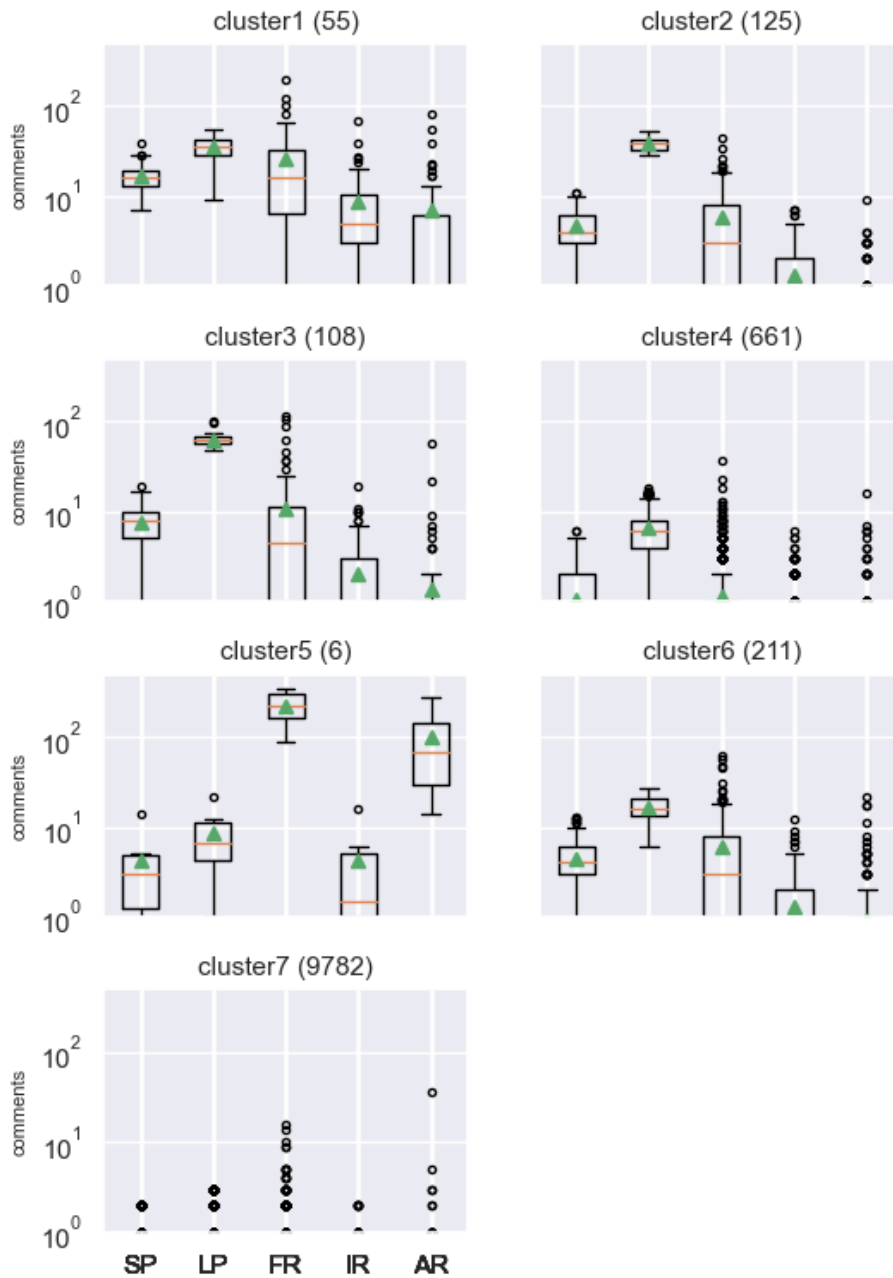


FIGURE L.12: Distribution of dialogic features in clusters found by the X-Means algorithm on the fifth run of Understanding Language, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows: **cluster1** : 8bb-more active SL who do not give additional replies; **cluster2** : 7-ASL without turn-taking; **cluster3** : 7a-more active SL without turn-taking; **cluster4** : 2-loners; **cluster5** : 8-active social learners; **cluster6** : 8b-ASL who do not give additional replies; **cluster7** : 1-asocial learners.

## X-means clustering on understanding-language (run 6)

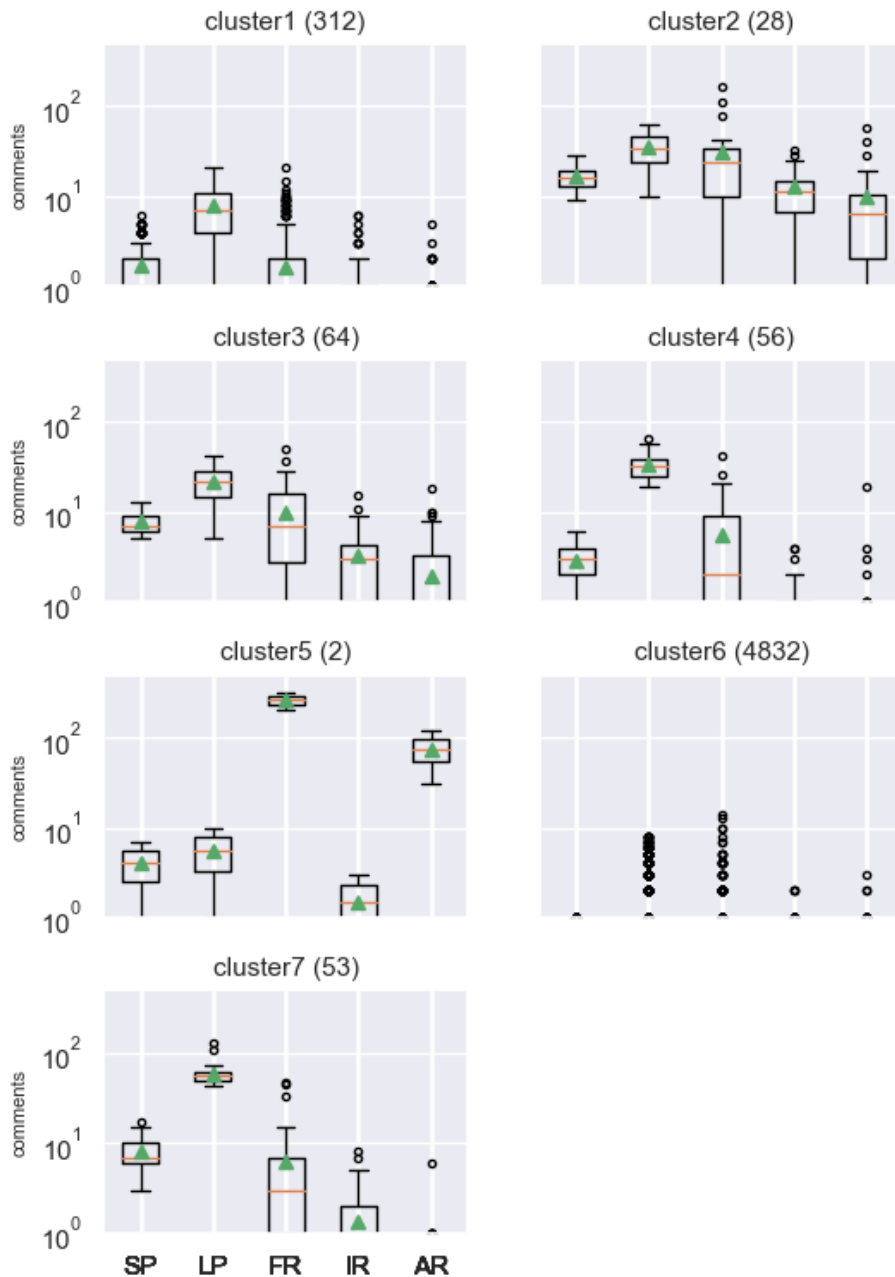


FIGURE L.13: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Understanding Language MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows: **cluster1** : 2-loners; **cluster2** : 8-active social learners; **cluster3** : 8b-ASL who do not give additional replies; **cluster4** : 7-ASL without turn-taking; **cluster5** : 8a-more active social learners; **cluster6** : 1-asocial learners; **cluster7** : 7a-more active SL without turn-taking.



## X-means clustering on understanding-language (run 7)

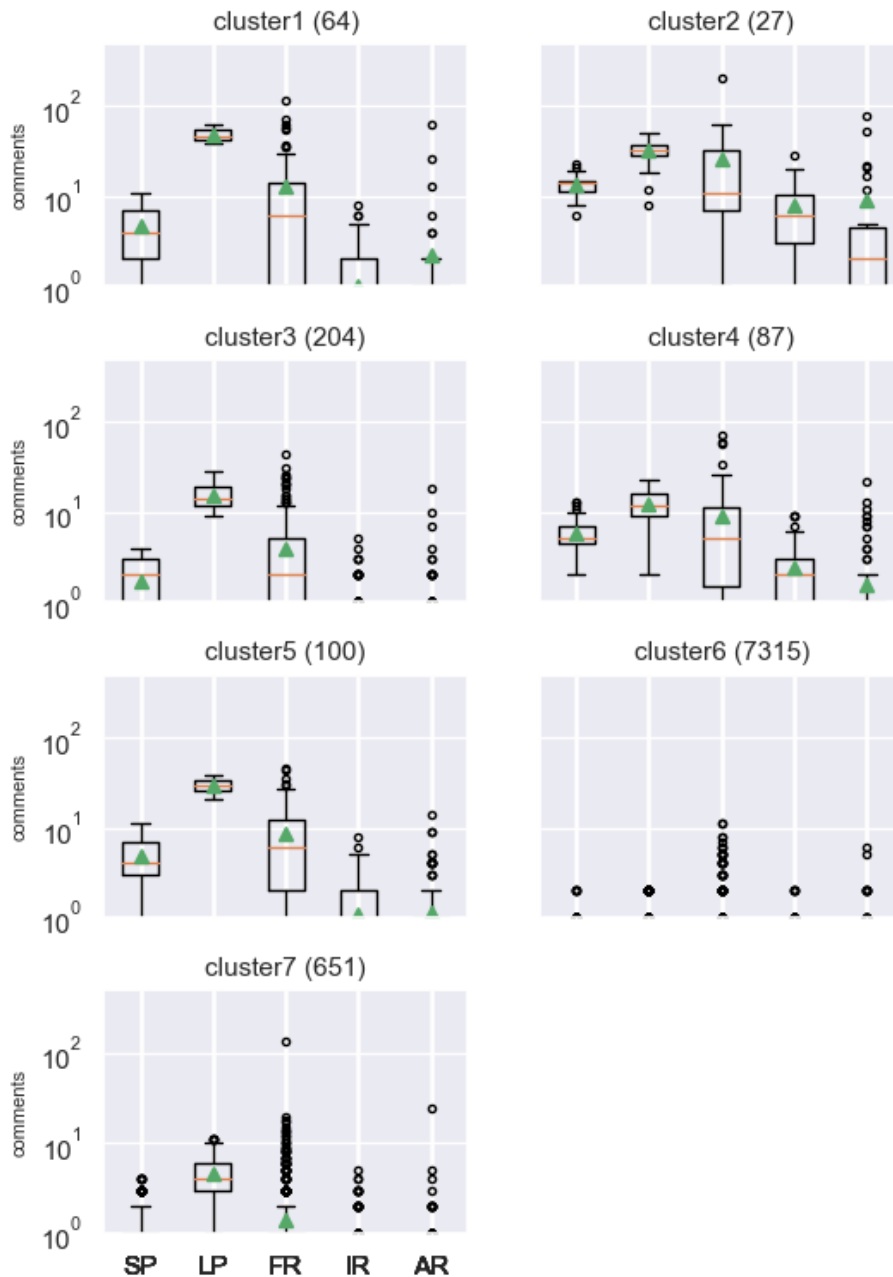


FIGURE L.14: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Understanding Language MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows: **cluster1**: 7a-more active SL without turn-taking; **cluster2**: 8a-more active social learners; **cluster3**: reluctant active social learners; **cluster4**: 8-active social learners; **cluster5**: 7-ASL without turn-taking; **cluster6**: 1-asocial learners; **cluster7**: 2-loners.

## X-means clustering on understanding-language (run 8)

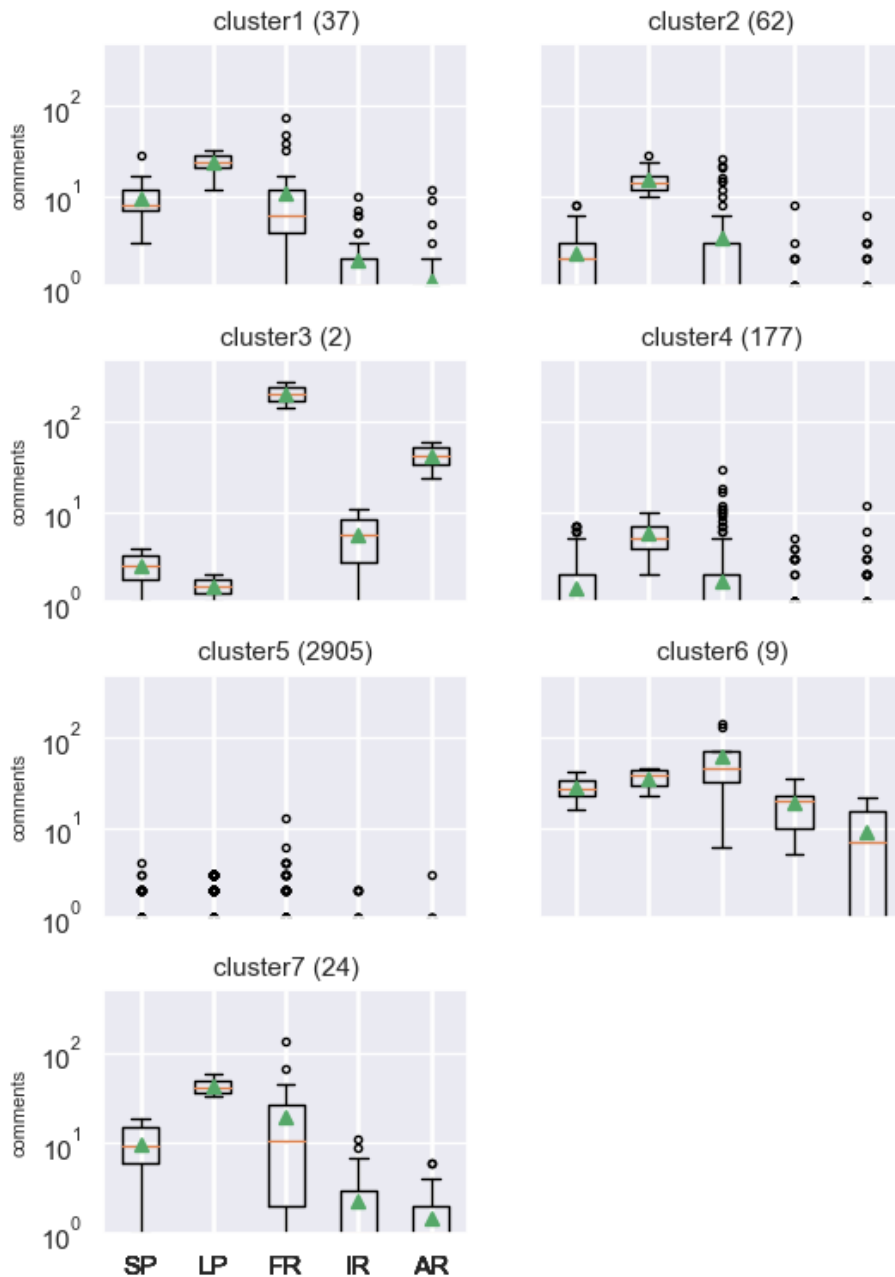


FIGURE L.15: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Understanding Language MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1** : 7-ASL without turn-taking; **cluster2** : less active social learners; **cluster3** : 8a-more active social learners; **cluster4** : 2-loners; **cluster5** : 1-asocial learners; **cluster6** : 8-active social learners; **cluster7** : 7a-more active SL without turn-taking.

## X-means clustering on understanding-language (run 9)

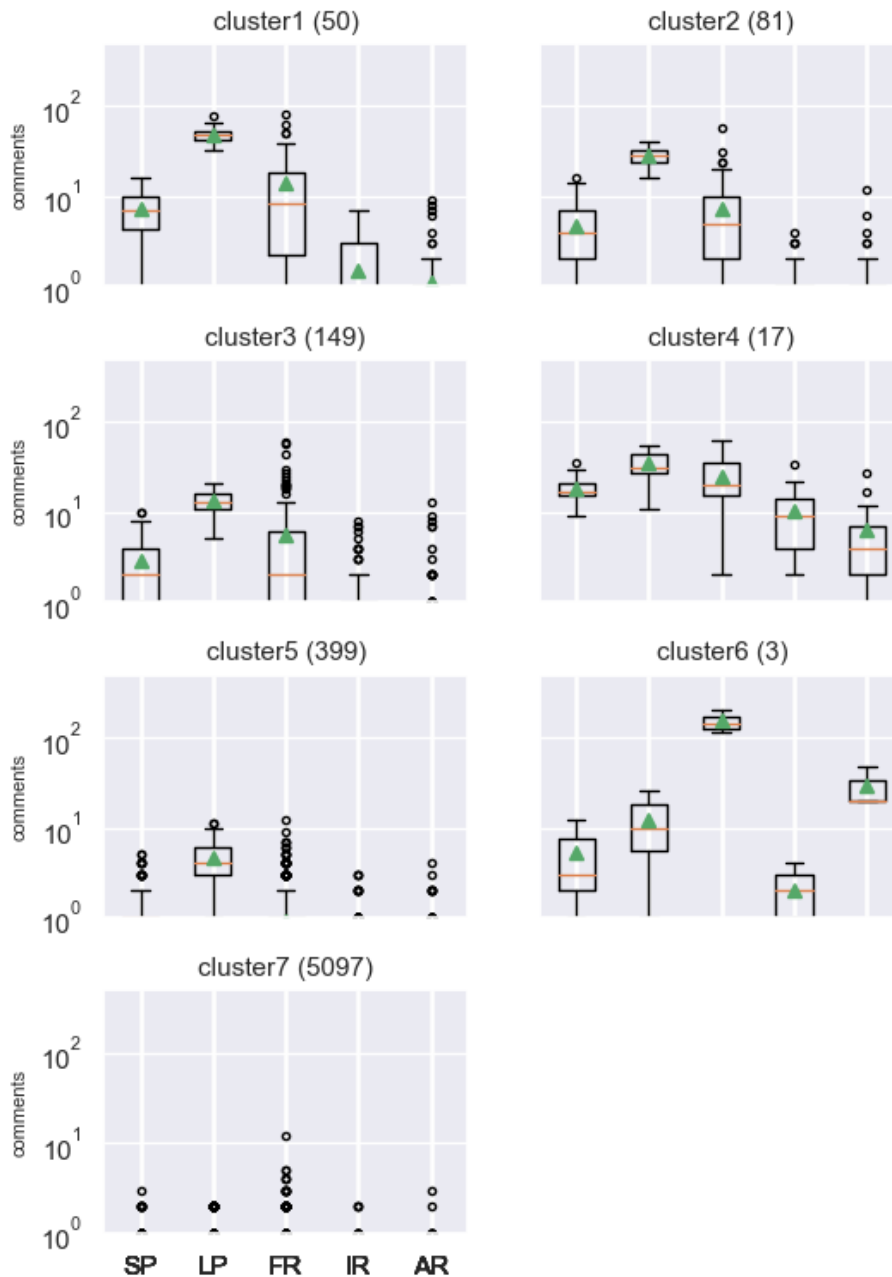


FIGURE L.16: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Understanding Language MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1** : 8b-ASL who do not give additional replies; **cluster2** : 7a-more active SL without turn-taking; **cluster3** : 7-ASL without turn-taking; **cluster4** : 8-active social learners; **cluster5** : 2-loners; **cluster6** : 8a-more active social learners; **cluster7** : 1-asocial learners.

## X-means clustering on understanding-language (run 10)

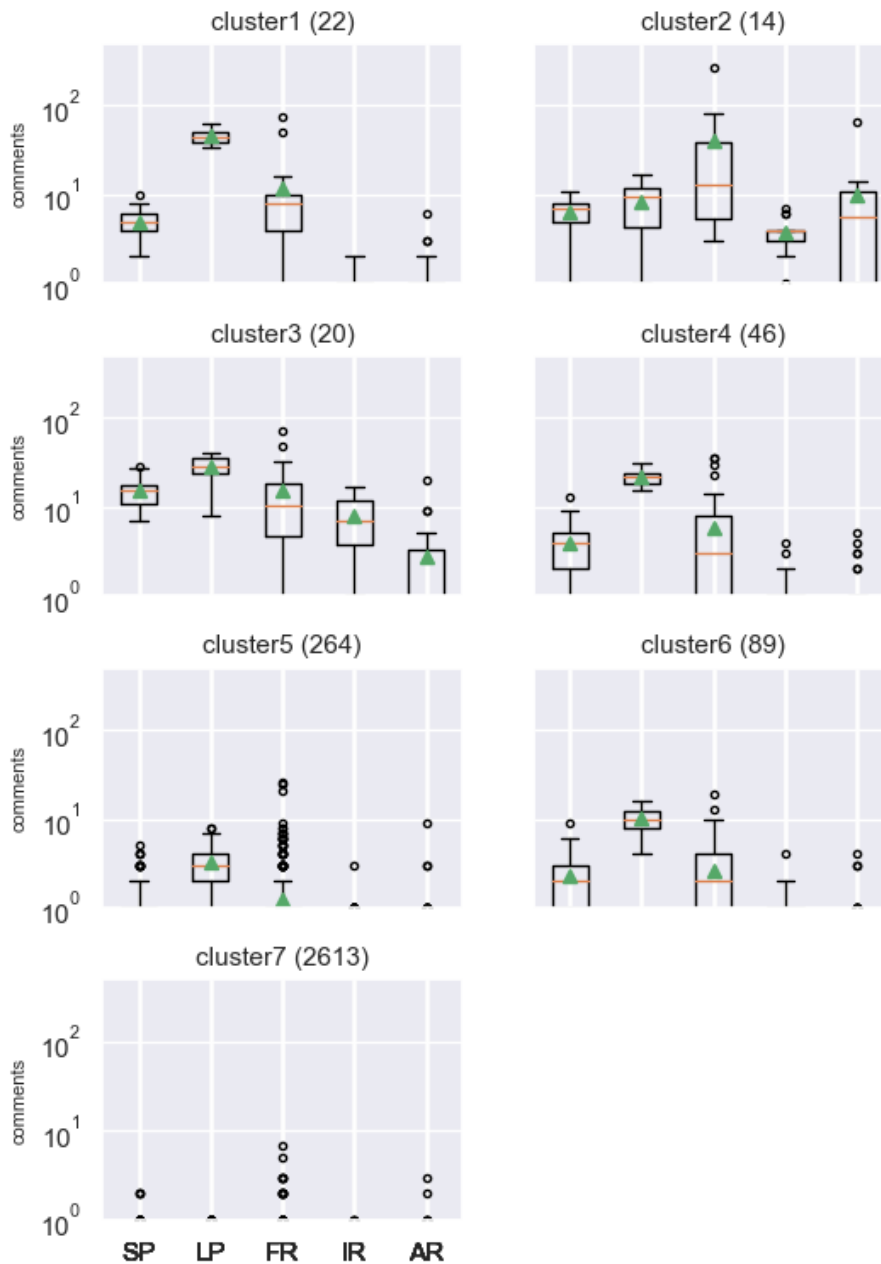


FIGURE L.17: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Understanding Language MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows. **cluster1**: 7aa-even more active SL without turn-taking; **cluster2**: 8b-ASL who do not give additional replies; **cluster3**: 8-active social learners; **cluster4**: 7a-more active SL without turn-taking; **cluster5**: 2-loners; **cluster6**: 7-ASL without turn-taking; **cluster7**: 1-asocial learners.

## X-means clustering on understanding-language (run 11)

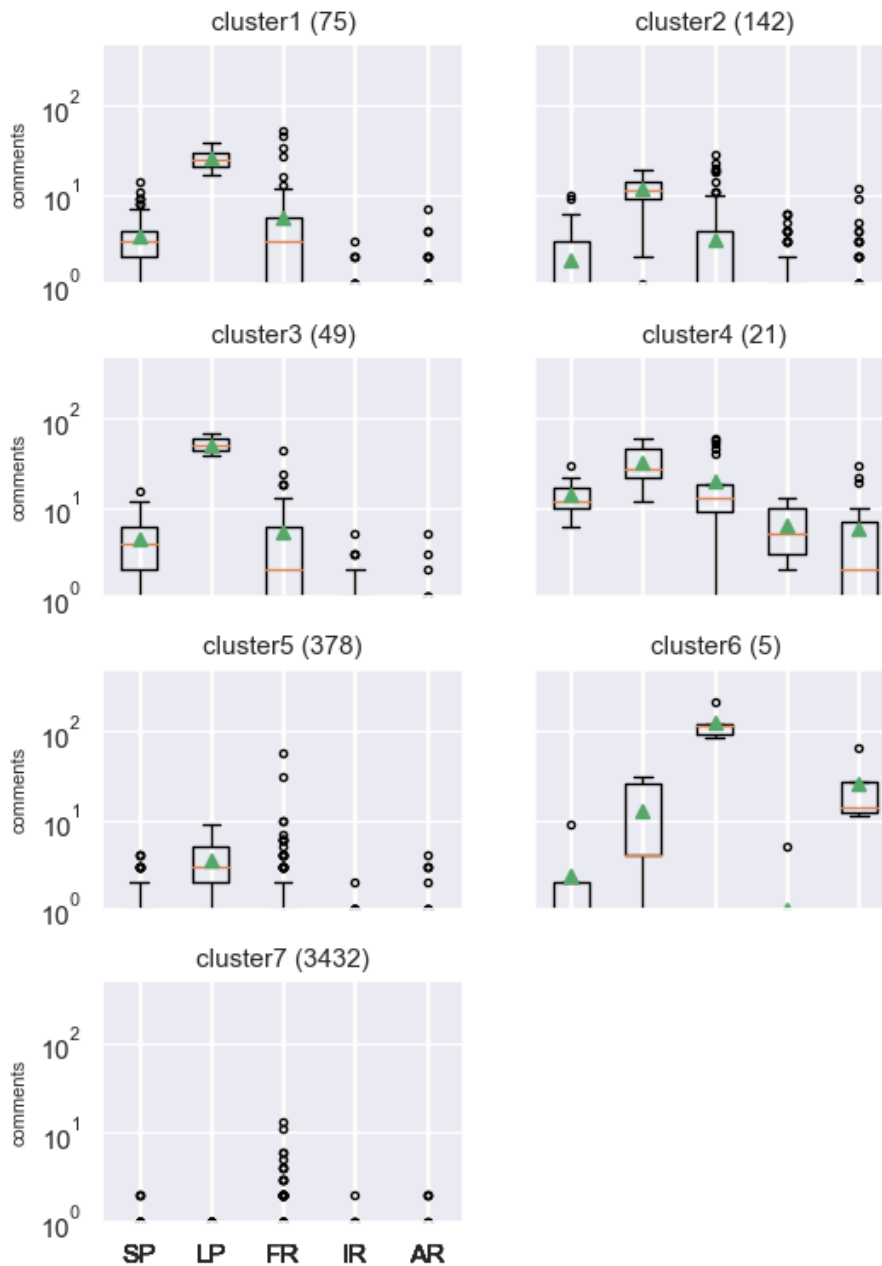


FIGURE L.18: Distribution of dialogic features in clusters found by the X-Means algorithm on the first run of the Understanding Language MOOC, with  $k=7$ . The semantics for each cluster, based on the median values for the dialogic features are as follows.

**cluster1** : 7-ASL without turn-taking; **cluster2** : 2a-more active loners; **cluster3** : 7a-more active SL without turn-taking; **cluster4** : 8-active social learners; **cluster5** : 2-loners; **cluster6** : 6-reluctant ASL; **cluster7** : 1-asocial learners.

