UNIVERSITY OF CANTERBURY

Supporting Engagement in Active Video Watching Using Quality Nudges and Visualisations

A thesis submitted in partial fulfilment of the requirements for the Degree of Doctor of Philosophy in Computer Science

by Negar Mohammadhassan

Supervised by:

Professor Antonija Mitrovic

Dr. Kourosh Neshatian

Dr. Jonathan Dunn

JUNE 1, 2022

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I am extremely grateful to my supervisor, Prof. Tanja Mitrovic, for her invaluable guidance, continuous support and dedicated involvement during my PhD study. Her wealth of knowledge and plentiful experience have been a constant inspiration in my academic research and daily life. My sincere thanks also go to my associate supervisors, Dr Kourosh Neshatian and Dr Jonathan Dunn, for their insightful feedback at every stage of this research project. I would like to thank Prof. Vania Dimitrova for her technical feedback on my study. I would like to express my special gratitude to Jay Holland for his treasured technical support as well as his wonderful company in the Intelligent Computer Tutoring Group (ICTG). I would like to extend my gratitude to all members of the ICTG lab: Jaafaru Musa for his effective collaboration, Faiza Tahir for bringing unconditional kindness to the lab, and Geela Fabic for encouraging me to start this PhD study. I am deeply grateful to the College of Engineering, and Computer Science and Software Engineering Department of the University of Canterbury for their financial support and providing a friendly yet professional environment for learning and growing.

I cannot find words to express my gratitude to my parents and siblings for their unfailing support, love and attention throughout my study. Although COVID19 travel restrictions in the last two years kept me away from my family, they were always available to talk to me and boost my confidence and perseverance. Finally, I would like to thank my partner for his tremendous understanding and constant encouragement. This accomplishment would not have been possible without my family and partner. Thank you.

Abstract

Learning by watching videos has been a popular method in e-learning. However, developing and maintaining constructive engagement is a crucial challenge in video-based learning (VBL). AVW-Space is an online VBL platform that enhances student engagement by providing note-taking and peer-reviewing. Previous studies with AVW-Space showed higher learning outcomes for students who write high-quality comments. Furthermore, an earlier study on AVW-Space suggested that visualising the student progress could help learners monitor and regulate their learning. Thus, this research aimed to increase engagement in AVW-Space by offering 1) personalised prompts, named Quality nudges, to encourage writing better comments and 2) visualisations of the student model to facilitate monitoring and controlling learning.

I conducted a series of studies to investigate the effectiveness of Quality nudges and visualisations on the students' engagement and learning. Firstly, I automated the assessment of comments quality using machine learning approaches. Then, I developed Quality nudges which encourage students to write better comments by triggering critical thinking and self-reflection. Next, I conducted a study in the context of presentation skills to analyse the effectiveness of the Quality nudges. The results showed that Quality nudges improved the quality of comments and increased learning consequently. After adding new visual learning analytics to AVW-Space, I investigated the effectiveness of the visualisations by conducting another study in the context of presentation skills. The results showed that the visualisations enhanced constructive engagement and learning even further. I also investigated the generalisability of nudges and visualisation for another transferable skill by making Quality nudges and visualisations customisable and conducting a study in the context of communication skills. Although the results showed that students used visualisations and nudges for communication skills differently from the participants in the study on presentation skills, findings indicated these interventions were still effective in increasing the quality of comments and enhancing constructive behaviour and learning.

This research contributes to the development of intelligent learning environments which provide personalised interventions to encourage constructive commenting behaviours during video-based learning. The interventions proposed in this research can be applied to other domains which involve critical thinking and self-reflection. Another contribution of this research is providing visual learning analytics for students in VBL platforms to increase learning awareness and engagement. The nudges and visualisations proposed in this research could be applied to any other video-based learning platform that allows commenting.

Deputy Vice-Chancellor's Office
Postgraduate Research Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapters 3 and 4 are the expanded versions of publications listed in Section 1.5

Please detail the nature and extent (%) of contribution by the candidate:

70%

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the Doctoral candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Tanja Mitrovic* Date: *20/1/2022*

Glossary

AIC: Akaike's Information Criterion

ANOVA: Analysis of Variance

API: Application Programming Interfaces

AVW-Space: Active Video Watching-Space

BCW: Behaviour Change Wheel

CFI: Comparative Fit Index

CI: Confidence Interval

CK1/2: pre-/post- study Conceptual Knowledge

COMB: Capability, Opportunity and Motivation Behaviour model

EFL: English as a Foreign Language

ENA: Epistemic Network Analysis

GLM: Generalised Linear Model

ICAP: Interactive, Constructive, Active, Passive

LIWC: Linguistic Inquiry and Word Count

ML: Machine Learning

MOOC: Massive Open Online Course

MSLQ: Motivated Strategies for Learning Questionnaire

NASA-TLX: NASA Task Load Index

NLP: Natural Language Processing

NTA: Network Text Analysis

PG: Postgraduate

QE: Quantitative Ethnography

QN: Quality nudges

RMSEA: Root Mean Square Error of Approximation

RN: Reminder nudges

SD: Standard deviation

SRL: Self-regulated Learning

SVM: Support Vector Machine

TAM: Technology Acceptance Model

VBL: Video-based Learning

VLA: Visual Learning Analytics

1 Introduction

1.1 Motivation

Videos have been a popular educational medium in e-learning. New technologies such as smartphones and tablets in combination with social media such as YouTube have facilitated the integration of video applications in education (Ebied et al., 2016; Snelson et al., 2012). Furthermore, learning from videos has increased remarkably due to the worldwide closure of educational institutions during the COVID19 pandemic (UNESCO, 2020). Video-based learning (VBL) is used in formal learning platforms such as Massive Open Online Courses (MOOCs) or informal learning media such as YouTube. This type of learning is also utilised in the flipped classroom, where the students are first asked to watch instructional videos and then discuss the concepts or perform further relevant educational activities in the classroom (Gilboy et al., 2015). Therefore, leveraging VBL in classrooms enables teachers to spend more time scaffolding the concepts using in-class activities rather than just instructing (Tucker, 2012). VBL can also be used in the form of self-assessment, in which students watch recorded videos of their performance to reflect on their learning progress and strategies (Colasante & Colasante, 2011; Hulsman & van der Vloodt, 2015).

Videos can simplify complicated concepts for learners by combining texts, visual effects, animations and sound (Ahmet et al., 2018; Sherin & Es, 2009). In addition, VBL can motivate learners to increase their engagement (Steffes & Duverger, 2012), resulting in better learning outcomes (Boateng et al., 2016; Mendoza et al., 2015). VBL also allows learners to learn at their own pace and anywhere (Gilboy et al., 2015). However, the lack of interaction between students and teachers in VBL might result in passive learning, in which the student is merely a receiver of information (Alavi, 1994; Yousef et al., 2014). Therefore, VBL platforms need to provide interactivity and interventions to support constructive learning (Giannakos et al., 2016; Ulrich et al., 2021; D. Zhang et al., 2006).

There have been several approaches to increase student engagement in VBL, such as embedding annotation tools, quizzes, learning resource recommendations and visual learning analytics (Chatti et al., 2016; Giannakos et al., 2016; Schulten et al., 2020; Wachtler et al., 2016). There are also collaborative tools integrated into VBL, such as forums (Chen et al., 2019), to increase engagement by social interaction and knowledge sharing. However, those tools do not provide adaptation and personalised intervention or feedback to learners. Thus, (Giannakos et al., 2016) pointed to the need for smart VBL environments, which would personalise interactions to support engagement for a wide range of learners with different levels of background knowledge and interests. Hence, one of the most important research topics in online learning is providing personalised support for learners engagement in VBL (Chatti et al., 2016; Giannakos et al., 2016; Mirriahi et al., 2021).

Active Video Watching (AVW)-Space (Mitrovic et al., 2016; Mitrovic, Dimitrova, et al., 2017) is an online VBL platform that was developed at the University of Canterbury initially for teaching transferable skills (e.g., presentation and communication). However, the platform can be used for teaching other types of skills. AVW-Space aims at facilitating engagement in VBL via a variety of approaches such as note-taking, micro-scaffolding (aspects), reviewing, visualisations and personalised prompts (Dimitrova et al., 2017; Mitrovic et al., 2019). In AVW-Space, the teacher first selects instructional videos from YouTube. Next, the teacher should define some aspects as micro-scaffolds to direct students' attention to the key concepts or encourage them to reflect on their relevant experience or knowledge. Then, learning happens in two main phases in AVW-Space: Personal Space and Social Space. In Personal Space, the learner can watch a video, pause it to write a comment and tag the comment with one of the aspects the teacher has defined. Once the teacher chooses the comments to be peer-reviewed by students anonymously, the Social Space becomes accessible to the students. Next, the student can review the comments made by their classmates and rate them using the rating options defined by the teacher in the Social Space.

An early study with AVW-Space (Mitrovic, Dimitrova, et al., 2017) in the context of presentation skills found that students who commented on videos learned more than their peers who watched videos passively. This study also showed that the use of aspects and rating options had a positive effect on learning. Therefore, prompts and signposting were suggested to enhance AVW-Space in supporting engagement and constructive learning. This study also emphasised the need for developing a detailed learner model to provide personalised prompts and signposting. This learner model should incorporate explicit profiling (e.g., asking students about their experience in the target skill and their learning strategies) and implicit profiling from the interaction logs (e.g., the number of comments made, videos watched, use of aspects and ratings).

AVW-Space introduced intelligent prompts called nudges, which are personalised interventions aimed to influence user behaviour towards constructive learning without limiting students' personal choices for engaging in AVW-Space (Mitrovic, Dimitrova, et al., 2017). The initial set of nudges implemented in AVW-Space were the Reminder nudges (RN) to encourage students to write comments and use various aspects (Mitrovic et al., 2019). An evaluation study showed that the Reminder nudges resulted in more comments, better usage of aspects, and increased learning (Mitrovic et al., 2019). However, the analysis of the data collected from previous studies with AVW-Space showed that students who wrote high-quality comments (i.e., comments in which students reflected on their experience or planned for their future presentations) learned more than students who made comments merely about the video (Taskin et al., 2019). The revealed correlation between high-quality comments and learning highlighted the need for more sophisticated nudges to encourage students to write high-quality comments.

The previous studies on AVW-Space also indicated that students use different learning strategies and require different supports to understand the task values and regulate their learning progress (Dimitrova et al., 2017; Mitrovic, Dimitrova, et al., 2017). Thus, signposting in the form of visual learning analytics of the student and social models was suggested to facilitate students' awareness of their engagement and increase motivation (Aguilar et al., 2021; Hooshyar et al., 2020; Mitrovic, Dimitrova, et al., 2017). An early version of AVW-Space provided a visualisation of highly attended video parts, which received more comments from previous learners. A previous study on AVW-Space showed that adding Reminder nudges and the visual learning analytics enhanced engagement and reduced frustration in commenting (Mitrovic et al., 2019). However, the visualisations of important parts of the videos do not give the learners insight into their own progress to regulate their learning. Thus, more provoking and informative visualisations of the student performance are required to allow students to monitor their progress and control their learning.

To summarise, AVW-Space demands more personalised and granular nudges as well as visual learning analytics to provide more specific support for improving engagement and fostering constructive VBL.

1.2 Research Plan

The research presented in this thesis aims to increase engagement in AVW-Space by 1) developing personalised prompts for improving the quality of comments, and 2) providing evocative visual analytics of the student model. In the first phase of this research, I extend the standard version of AVW-Space to assess the quality of comments as students write them and provide Quality nudges that provoke reflection and encourage writing high-quality comments. Then, I investigate the effectiveness of Quality nudges in increasing engagement and learning outcomes in AVW-Space. In the second phase of this research, I add new visual learning analytics to AVW-Space to help the learners identify the learning tasks they need to complete and monitor their own learning performance. Next, I analyse the effectiveness of the visualisations in increasing engagement and learning in AVW-Space. I limit the focus of the first two phases only to the context of presentation skills. However, in the third phase of this research, I customise Quality nudges and visual learning analytics for another domain in AVW-Space. Then, I investigate the effectiveness of nudges and visualisations in the context of a new transferable skill. The following section details these three phases of my research, their scope and the research questions to be addressed in them.

1.3 Scope of the Project and Research Questions

The scope of this research is increasing engagement and learning in the AVW-Space platform. However, the methodology taken in this project is applicable to other VBL platforms. It is important to note that the quality of videos is an essential factor in students' engagement (Mayer, 2021). This research uses the videos carefully selected for previous AVW-Space studies, but investigating the effects of video quality is not the focus of this research. In addition, the participants in my studies are students at the University of Canterbury. I also narrow the context of using AVW-Space to transferable skills for this PhD research.

In the first phase of my research, I propose and evaluate the Quality nudges by addressing the following research questions:

- *RQ1. How can I define a robust scheme for human coders to assess the quality of comments?*
- *RQ2. How reliable are the machine learning models trained using the quality scheme to classify comments based on their quality?*
- *RQ3. How can I design and develop Quality nudges to encourage students to improve the quality of their comments?*
- *RQ4. Do Quality nudges increase engagement and learning for students?*

I address RQ1 by proposing a quality scheme for assessing the quality of comments and evaluating the human coders' agreement on this quality scheme. Next, RQ2 is addressed by presenting the development and evaluation of machine learning classifiers that assess the quality of comments based on the quality scheme. Then, the design and development of Quality nudges are presented to address RQ3. To address RQ4, I conduct the first study with quasi-experimental design on a large, first-year engineering course at the University of Canterbury. In this study, students use a version of AVW-Space that includes Quality nudges in addition to Reminder nudges. A quasi-experimental design is used to estimate the causal impact of an intervention on the target population without any assignment into treatment or control groups. Instead, the results of a quasi-experiment with the new intervention are compared to prior studies without that intervention.

In the second phase of my research, I leverage the findings from the previous studies to design new useful visual learning analytics for AVW-Space. After implementing these visualisations, I present the second quasi-experimental study on students enrolled in the same engineering course to evaluate the effectiveness of new visual learning analytics. The version of AVW-Space used in this study contains the visual learning analytics, on top of the Quality and Reminder nudges. The data collected from the second evaluation study is analysed to investigate the effectiveness of visualisations and address the following research question:

RQ5. Does the visualisation of the student model increase engagement and learning?

Since the first two phases of my research are in the context of presentation skills, the third phase of my research investigates the generalisability of these enhancements (Quality nudges and visual learning analytics), and address the following research questions:

RQ6. How can the quality assessment models, nudges and visual learning analytics be generalised to other soft skills?

RQ7. Do nudges and visual learning analytics increase engagement and learning in other soft skills?

I address RQ6 by explaining the development of customisable nudges and visual learning analytics. Then, I conduct the third quasi-experimental study in a second-year software engineering course in the context of learning face-to-face team meeting communication. In this study, participants use a version of AVW-Space that includes Reminder and Quality nudges as well as the visual analytics customised to the domain. The objective of this study is to see if the findings on the effectiveness of the nudges and visual learning analytics in the context of presentation skills could be generalised to other transferable skills. Finally, RQ7 is discussed by analysing data collected from the third evaluation study.

1.4 Contributions

One of the essential contributions of this research is the design and evaluation of a quality scheme for assessing the quality of comments. The presented quality scheme can also be used in other video-based learning systems that provide annotation. Another main contribution is the development of intelligent and personalised nudges that encourage students to make high-quality comments. Moreover, this research evaluates the effectiveness of integrating visual learning analytics into VBL to facilitate learners' awareness of their learning progress and help them regulate their learning. Most importantly, this research investigates the feasibility and effectiveness of generalising the personalised nudges and visual learning analytics for different soft skills. Additionally, this research presents behavioural patterns which lead to constructive learning in VBL. The result from mining behavioural patterns provides insights into potential supports for engagement in VBL for future work.

1.5 List of Research Publications

Journal articles:

- Mohammadhassan, N., Mitrovic, A., Neshatian, K. (2022) Investigating the Effect of Nudges for Improving Comment Quality in Active Video Watching. *Computers & Education*, 176, 104340 (open access) https://doi.org/10.1016/j.compedu.2021.104340

Conference proceedings:

- Mohammadhassan, N., Mitrovic, A. (2022) Investigating the Effectiveness of Visual Learning Analytics in Active Video Watching, The 23rd International Conference on Artificial Intelligence in Education (AIED 2022) [in print].

- Mohammadhassan, N., Mitrovic, A. (2022) Discovering Differences in Learning Behaviors during Active Video Watching using Epistemic Network Analysis, In: Wasson B., Zörgő S. (eds) Advances in Quantitative Ethnography. ICQE 2021. Communications in Computer and Information Science, vol 1522. Springer, Cham. https://doi.org/10.1007/978-3-030-93859-8_24
- Mohammadhassan, N., Mitrovic, A. (2021) Investigating Engagement and Learning Differences between Native and EFL students in Active Video Watching, In: Rodrigo, M. M. T. et al. (Eds.) Proceedings of the 29th International Conference on Computers in Education, pp. 1-10. Asia-Pacific Society for Computers in Education.
- Mohammadhassan, N., Mitrovic, A., Neshatian, K., Dunn, J. (2020) Automatic Assessment of Comment Quality in Active Video Watching. In: So, H.J. et al. (Eds.) Proceedings of the 28th International Conference on Computers in Education, pp. 1-10. Asia-Pacific Society for Computers in Education.

Doctoral track paper and demo:

- Mohammadhassan, N., Mitrovic, A. (2021) Providing Personalized Nudges for Improving Comments Quality in Active Video Watching, Companion Proceedings of the 11th International Conference on Learning Analytics & Knowledge LAK 2021, pp. 145 (demo)
- Mohammadhassan, N., Mitrovic, A., Neshatian, K., Dunn, J. (2020) Developing Personalized Nudges to Improve Quality of Comments in Active Video Watching. In: So, H.J. et al. (Eds.) Proceedings of the 28th International Conference on Computers in Education, pp. 766-769. Asia-Pacific Society for Computers in Education. (Doctoral Student Consortium)

1.6 Thesis Structure

Chapter 1 delivers the motivation, goals, and contributions of this research. Chapter 2 reviews the literature on engagement difficulties in VBL as well as approaches taken to enhance engagement in VBL, prior work in AVW-Space and Epistemic Network Analysis for understanding students' behaviours. Chapter 3 presents the design and evaluation of the proposed quality assessment scheme and classifier, followed by the design and implementation of Quality nudges. Some material covered in Chapter 3 has been previously published in (Mohammadhassan et al., 2020, 2022). Chapters 4-6 cover the first, second and third evaluation studies, respectively. Chapter 4 material has been previously published in (Mohammadhassan et al., 2022; Mohammadhassan & Mitrovic, 2021a, 2022a) and demonstrated in (Mohammadhassan & Mitrovic, 2021b). Some results presented in Chapter 5 have been submitted and accepted in (Mohammadhassan & Mitrovic, 2022b). Finally, the last chapter (Chapter 7) binds the entire research project and proposes research directions for the future.

2 Literature Review

The primary challenge in video-based learning is establishing and maintaining student engagement. In education, engagement contains three main interconnected dimensions (Fredricks et al., 2004): behavioural engagement (students' performance in learning activities), emotional engagement (students' positive and negative reactions to teachers and classmates) and cognitive engagement (students' level of investment in learning, or being strategic). Simply delivering educational videos is insufficient to foster cognitive engagement. The lack of human interaction, interactivity with the video, effective feedback and personalisation can cause video-based learning to be perceived as a passive form of learning (Chatti et al., 2016; Yousef et al., 2014). Therefore, video-based learning requires support that encourages students to engage with the content conscientiously.

There have been a variety of strategies to increase student engagement in video-based learning, including embedding quizzes, annotation tools, learning resource recommendations, collaboration tools and visualisations (Chatti et al., 2016; Giannakos et al., 2016; Schulten et al., 2020; Wachtler et al., 2016). For creating interactive VBL environments, teachers can use open source and commercial technologies to include quizzes, visualisation, branching to various video segments, and learning analytics (Kleftodimos & Evangelidis, 2016). However, this PhD research develops its own set of engagement-supporting tools in order to offer personalisation. Towards this goal, this research investigates quantitative data (e.g., the number of interactions, time spent on the system and pre-/post-test scores) along with qualitative data (e.g., interaction logs) using statistical analysis and quantitative ethnography to deeply understand students' behaviour.

Section 2.1 provides an overview of various approaches for increasing engagement in video-based learning and discusses their effectiveness. Section 2.2 presents previous work on AVW-Space to underline where AVW-Space falls short in terms of engagement supports. Section 2.3 presents a brief description of quantitative ethnography and its most common tool, Epistemic Network Analysis (ENA), used in this research. Finally, Section 2.4 summarises the gaps in the literature which this PhD research aims to address.

2.1 Engagement Strategies in Video-Based Learning

This section provides an overview of previous work on the most common approaches for supporting engagement in video-based learning (i.e., quizzes, prompts, annotation tools, visualisations, collaborative tools and visualisations) and highlights their limitations.

2.1.1 Quizzes and Prompts

A number of studies have investigated the effectiveness of using quizzes within educational videos. Cummins et al. (2016) demonstrated embedding quizzes in programming video lectures which offered instant feedback from the teaching staff on the learners' answers. Although these quizzes effectively increased engagement, giving personalised feedback could be impractical in a large class. Haagsman et al. (2020) examined the effectiveness of in-video quizzes that offered automatic and on-demand feedback by displaying the relevant video segment. The study discovered that the in-video quizzes promoted active viewing behaviour. However, asking questions about specific parts of the video may lead the student to concentrate exclusively on those parts to merely answer the question rather than learn the video's content. To avoid this undesirable behaviour, Mirriahi et al. (2021) curated a set of questions that capture the student's ability to combine and apply learned information into other contexts. The evaluation study showed that these pre-defined quizzes with immediate feedback promoted active learning and self-efficacy. Despite the effectiveness of quizzes, Shelton et al. (2016) showed that in-video quizzes could be distracting. In addition, Rice et al. (2019) suggested designers should be mindful that quizzes could be anxiety-inducing. Moreover, the nature of quizzes focuses on specific information and restricts students from forming their ideas. Although quizzes are transferable and can be used over several years, designing and developing in-video quizzes could be time-consuming and demanding for instructors.

Prompts are explicit interventions that can be added to videos to enhance learners' engagement. In contrast to quizzes, prompts are more generic and allow more flexibility to students in forming their understanding. Several studies have been conducted on in-video prompts to support meta-cognitive activities necessary for learning, such as self-reflection and self-regulation (Bannert & Mengelkamp, 2008). Self-reflection entails judging and evaluating used strategies (Zimmerman, 2000). Self-regulation is the process through which self-generated thoughts, feelings, and actions are planned and adapted to accomplish personal goals (Zimmerman, 2000). Shin et al. (2018) investigated how learners and instructors in the higher education context perceive in-videos prompts in which students respond to reflective questions while watching videos. They discovered that although some learners considered prompts as beneficial reflection checkpoints, others found them distracting. Additionally, they found that different prompting formats suited different learners, which highlights the need for personalisation. Alten et al. (2020) also found positive learning outcomes after adding self-regulated learning (SRL) prompts, which require eight-grade students to think about and answer them to continue watching the video. Although the SRL prompts assisted students to be more aware of their learning, some students disliked the SRL instruction. This dissatisfaction could be attributed to the fact that these prompts are content-based and not adapted to the learner's performance and behaviour. That results in high-performing learners receiving the same prompts as low-achieving students. Hence, prompts must be tailored to the learners' context and preferences.

Various computer-based learning platforms have integrated personalised prompts to support meta-cognitive skills and engagement. For example, an Intelligent Tutoring System for teaching English grammar offered adaptive prompts to encourage self-explanation of worked grammar examples (Wylie et al., 2011). MetaTutor (Azevedo et al., 2012; Bouchet et al., 2016) is a multi-agent learning environment that assists students in developing self-regulated learning using adaptive prompts. However, students answer the prompts using pre-defined choices. A web-based platform for reporting science projects utilises Natural Language Processing (NLP) to offer adaptive prompts that encourage students to revise their explanations by strengthening their evidence (Tansomboon et al., 2017). Another new Intelligent Tutoring System for Data Science also applies machine learning and NLP to provide personalised hints and Wikipedia-based explanations on students' written explanations (Kochmar et al., 2020). However, no personalised prompts have been implemented for video-based learning platforms. Thus, one of the objectives of this research is to introduce personalised prompts that trigger self-reflection and self-regulation to enhance engagement in video-based learning.

2.1.2 Annotation Tools

Annotation tools are another type of engagement support in video-based learning, which has been attracting interest in recent years due to improvements in web development technologies (Evi-Colombo et al., 2020). This approach is less demanding for teachers and allows learners to reflect on videos freely. Studies show that annotations have a beneficial influence on learning, attention and engagement (Chin-Yuan Lai et al., 2020; Pardo et al., 2015). The video annotation tools allow students to flag a particular video timestamp to reflect or review later (Dawson et al., 2012; Evi-Colombo et al., 2020). Sharing these annotations with the class could also allow students to gauge their personal learning in relation to others (Dawson et al., 2012). Some VBL platforms combine annotation tools with content highlighting. For instance, ViDex (Dodson et al., 2018) enables learners to add textual notes to videos and highlight intervals of videos or their transcripts. Chiu et al. (2018) present a VBL platform where students can annotate videos by highlighting a part of the video and commenting. C. Liu et al. (2019) introduced a note-taking platform in which learners first add notes on video transcripts and then reinterpret and synthesise their notes. Yoon et al. (2021) clustered students based on various interactions with a VBL platform and discovered that students who made annotations exhibited higher learning achievement than students who only browsed videos and their peers' annotations. However, some studies showed that the effectiveness of annotations in VBL is contingent upon the learner's learning strategies and motivation, emphasising the need of designing adaptive interventions for annotation (Mirriahi et al., 2016; Pardo et al., 2015). Therefore, recent studies leverage learning analytics and text mining to characterise the learning process in video annotations (Dodson et al., 2018; Joksimović et al., 2019; Seo et al., 2021).

Mining and applying learning analytics methods on students' interactions with the annotation tools could provide valuable insights into what parts of the video were important, confusing, or interesting for students (Dawson et al., 2012). An early study on edX MOOCs (Kim et al., 2014) combined collective and personal watching traces and labelled bookmarking of students to generate highlights of the videos as a summary for students reviewing. Another study derived various metrics from interactions with annotations (e.g., frequency, total number and timestamp) to investigate how learners develop learning strategies and engagement over time (Mirriahi, Jovanovic, et al., 2018). This study showed that some students sustain their engagement over time, some remain disengaged, and others fluctuate between low and high engagement levels (Mirriahi, Jovanovic, et al., 2018). This finding emphasises the importance of self-regulation and the need for personalised support in fostering engagement. However, these studies were only based on the count and timestamp of annotations and did not take their content into account.

The use of text analysis has been gaining momentum in VBL in order to detect misconceptions or poor engagement in the content of annotations. Recent research developed a taxonomy based on the referent types in video comments (referring to visual or verbal content or a concept in the video) (Yarmand et al., 2019). Then, these referent types were used to map comments to relevant video intervals to allow students to browse relevant comments made by peers in the adjacent discussion area (Yarmand et al., 2021). However, categorising comments based on the topic does not fully represent students' level of understanding and reflection. Hulsman et al. (2009) proposed four categories of reflection (Observations, Motives, Effects and Goals) to analyse students' annotations on the recorded video of their communication in DiViDu. This categorisation scheme inspired further research on the automated assessment of reflective notes (Joksimović et al., 2019; Mirriahi, Joksimović, et al., 2018). Gašević et al. (2014) investigated video annotations made by students in the CLAS note-taking environment (Risko et al., 2013) used Linguistic Inquiry and Word Count (LIWC) software (Tausczik & Pennebaker, 2010) to identify the linguistic characteristics of self-reflective annotations such as the length of words, pronouns, the tense of verbs, use of cognitive and perceptual vocabularies and emotional phrases. Following that, a scheme was suggested to specify the level of reflections in video annotations (Mirriahi, Joksimović, et al., 2018). The follow-up study utilised Coh-Matrix[1] (Graesser et al., 2014) to extract the linguistic properties of each level of reflection and evaluate the depth of students' self-reflection (Joksimović et al., 2019). However, these insights have not been used to provide automated personalised support for video-based annotations. One of the objectives of this PhD research is to address this gap.

2.1.3 Collaborative Tools

Accommodating tools for sharing knowledge and social interaction is another approach for increasing learning and engagement. It is crucial to provide a VBL environment that fosters collaborative knowledge creation while supporting the continuous creation of a personal knowledge network (Chatti et al., 2016). Vialogues (Agarwala et al., 2012) is an example of a VBL platform which provides a discussion area to enhance engagement and social learning. Similarly, (Chatti et al., 2016) introduced CourseMapper as a collaborative video annotation platform that allows students to share their knowledge about video lectures and vote on peers' annotations. CLAS (Risko et al., 2013) is another VBL environment that allows social bookmarking of the video parts. ConceptScape (C. Liu et al., 2018) also generates and presents a concept map for lecture videos by encouraging learners to collaboratively identify concepts of different video segments and their relationships. Discussion forums in MOOCs have also been effective in knowledge sharing (Almatrafi & Johri, 2019). A recent study showed that

[1] A computational linguistic facility for computing cohesion and coherence metrics of texts

posting on MOOC forums was an important indicator of success (Santos et al., 2014). However, the forums on MOOCs can contain broader topics (such as assignments, quizzes, etc.) than other collaborative tools, which focus on the video content.

The collaborative tools require teacher moderation and constructive collaboration of students. There have been supports suggested for addressing these requirements. For instance, the integration of personalised conversational agents into discussion forums has been proposed to provide feedback on learners' collaborative activities (Amarasinghe et al., 2019; Apoki & Crisan, 2019; Demetriadis et al., 2018). This PhD research offers intelligent and personalised support for active video watching, which is not directly focused on the collaborative tools but could improve students' constructive knowledge sharing and collaboration.

2.1.4 Visualisations

Visual learning analytics can provide insights into using learning resources (Chatti et al., 2016; Matcha et al., 2020) and the student's learning progress. The former is usually the same for all students, whereas the latter allows for greater adaptation and personalisation (Guerra et al., 2016). Visual analytics of the learner model is a learners' awareness tool that provides learners with up-to-date information on their learning status (Bodily et al., 2018). The learner model contains various metrics such as the learner's progress in learning activities, knowledge and affective states (Bull & Kay, 2010). Visualising this information assists learners in monitoring and assessing their learning and making informed plans to attain their learning objectives and control their learning (Bull & Kay, 2010; Matcha et al., 2020; Mitrovic & Martin, 2007). As a result, offering visualisations to students can boost engagement, motivation and learning (Aguilar et al., 2021; Hooshyar et al., 2020; Verbert et al., 2014). Additionally, the visualisation of the learning process can offer evocative insight, prompt self-reflection, and potentially provide interventions to optimise learning (Bodily et al., 2018; Muldner et al., 2015). In other words, visualisation of the student model can serve as feedback, providing visual cues to support students in evaluating their progress towards goals (Wang et al., 2011). However, the usefulness of visualisations depends on their explainability. Interpreting the visualisations and using the feedback presented to inform learning strategies could be difficult for learners (Corrin & de Barba, 2014). Besides, some visualisations may detract students' motivation and trigger social anxiety when they see the visual comparison of their peers' performance to their own (Lim et al., 2019; Lonn et al., 2015)

Visual learning analytics can present diverse types of information, the majority of which are cognitive and behavioural analytics (Matcha et al., 2020; Sedrakyan et al., 2020). Examples of cognitive visual analytics include competency tracking and displaying learning difficulties (Chou et al., 2017; Grann & Bushway, 2014; Mejia et al., 2017). Behavioural visual analytics could display the progression in learning tasks such as the number of solved problems or watched videos (Muldner et al., 2015; Rwitajit et al., 2019). Some visual learning analytics go beyond the domain knowledge to present the learner's metacognitive state, such as knowledge of study tactics and planning (Broos et al., 2017; Charleer et al., 2018; Santos et al., 2012) or students' emotional status to raise their emotional awareness (Ez-zaouia et al., 2020; S. Ruiz et al., 2016). On the other hand, some visualisations provide analytics of social models such as comparisons to the average or entire class and teamwork progress (Guerra et al., 2016; Vivian et al., 2015). Although the effectiveness of student-facing visual learning analytics has been researched in various computer-based educational platforms such as Learning Management Systems (Aguilar et al., 2014) and MOOCs (J. S. Ruiz et al., 2014), research on the application of visual analytics in video-based learning is still in its initial stages.

Visual analytics has been applied to the learners' interaction with video lectures, attitudes, and learning performance to find the most important part of the video (Risko et al., 2013), the difficult parts (Srivastava et al., 2019) and analyse class engagement across different segments of the video (Xia & Wilson, 2018). However, these visualisations were only for assisting teachers in decision making and were not displayed to the students. On the other hand, CourseMapper (Chatti et al., 2016) uses the log of students' interaction with video to provide a heatmap on the video scrub bar to help students identify the most viewed parts. CourseMapper also uses annotations timeframe and counts to display an annotation heatmap on the scrub bar. This heatmap illustrates video segments which received more annotations and likely contain interesting information. Similarly, ResponseCollector visualises the timeline of the students' responses (important, difficult, interesting and question) to the video (Okumoto et al., 2018). A study on edX MOOCs developed an automatically generated word cloud of concepts covered in the video intervals (Kim et al., 2014). However, these visualisations are the same for all students and do not provide any personalised information. Another type of visualisation in VBL is concept maps, where students link concepts of different video segments (C. Liu et al., 2018; S. Zhang et al., 2019). However, this visualisation is produced by students and do not provide any feedback on their learning or performance.

On the other hand, a video-based learning platform used in a flipped classroom (Yoon, Hill, et al., 2021) provides a simple visualisation of the student’s quiz scores and video completion rates to support students' self-monitoring and evaluation. An experiment with this visualisation showed that the learners who had access to this visualisation showed higher behavioural and cognitive engagement levels in pre-class (e.g., watching videos and answering the quizzes) and in-class sessions (e.g., team discussion, project work) without the instructor's reminders. However, this visualisation provided information on only the students' performance. The visualisation of the open learner model and open social model can make students aware of their behaviour and others behaviour and promote self-reflection and improve engagement (Brusilovsky et al., 2016). Thus, this PhD research proposes more detailed personal and social visualisation to address shortcomings in video-based learning and support engagement.

2.2 AVW-Space

Video-based learning is compelling for learning transferable skills (Conkey et al., 2013; Santucci et al., 2019; Kopolovich, 2020) since videos can emulate real experiences, allow contextualisation of personal experience and seeing various perspectives (Cronin & Cronin, 1992; Dimitrova & Mitrovic, 2021; Evi-Colombo et al., 2020). Learning transferable skills involves practising in different situations, receiving feedback, reflecting and practising again (Chadha, 2006; Sibthorp, 2003). However, offering such support to each student is time-consuming and resource-intensive for teachers (Anthony & Garner, 2016; Hetzner et al., 2011; Kopolovich, 2020). AVW-Space (Mitrovic et al., 2016) was initially developed as a controlled video-based learning platform for self-studying transferable skills, but it can be used in other domains. AVW-Space leverages the learners' familiarity with video watching and commenting on social media (e.g., YouTube). AVW-Space offers note-taking in the form of comments during video watching to facilitate student engagement and reflective learning. Furthermore, AVW-Space aims at developing a teacher-friendly environment. Teachers can select publicly available videos (e.g., YouTube videos) in AVW-Space instead of recording and editing their own videos. AVW-Space supports students' engagement via various approaches with minimal involvement from teachers. For a new learning space on AVW-Space, the teacher needs to choose YouTube videos and then identify the aspects for these videos. Aspects are micro-scaffolds aimed to direct the student's attention toward the fundamental concepts of the target skill or to trigger self-reflection. Once the teacher makes a new space available to the students, learning occurs in two phases: 1) Personal-Space, where students watch videos, write comments and tag them by aspects, and 2) Social Space, where students can review and rate comments anonymously using rating options defined by the teacher. Personal space is always available by students, while Social Space becomes accessible for learners once the teacher selects anonymised comments to be displayed for review.

Previous experimental studies on AVW-Space have been in the context of presentation skills (Mitrovic et al., 2016; Mitrovic, Dimitrova, et al., 2017; Mitrovic, Gostomski, et al., 2017; Mitrovic et al., 2019), which is also the main focus of this PhD research. In the space for presentation skills, there are four carefully selected tutorial videos for training students on giving oral presentations and four example videos of actual recorded oral presentations (Table 2-1). The aspects defined for tutorial videos contains three reflective aspects: *"I didn't realise I wasn't doing it"*, *"I am rather good at this"*, "*I did/saw this in the past*"; to trigger learners to reflect on their own experiences. The fifth aspect, "*I like this point*", allows learners to ascribe their learning points. The aspects for the example videos are "*Delivery*", "*Speech*", "*Structure*", and "*Visual aids*", which correspond to the concepts that tutorial videos covered. The rating options in Social Space are the second level of micro-scaffolds designed to prompt further reflection. These rating categories constitute: "*This is useful for me*", "*I hadn't thought of this*", "*I didn't notice this*", "*I don't agree with this*", and "*I like this point*".

Table 2-1 Description of videos used for oral presentation skills in AVW-Space.

Video	Title	Length (s)	YouTube ID
Tutorials			
1	How to Give an Awesome (PowerPoint) Presentation	174	i68a6M5FFBc
2	How to open and close presentations?	457	Yl_FJAOcFgQ
3	Make a presentation like Steve Jobs	415	RHX-xnP_G5s
4	The five secrets of speaking with confidence	382	7MWaeOHDBOg
Examples			
1	Abraham Heifets: How can we make better medicines? Computer tools for chemistry	203	0YdFyNZoTU0
2	Johanna Blakley: Social media and the end of gender	508	ZR4LdnFGzPk
3	Tim Berners-Lee: A Magna Carta for the web	408	rCplocVemjo
4	Jasdeep Saggar: Hypoxia-activated pro-drugs: a novel approach for breast cancer treatment	205	ZbkaJ7KnhXk

Figure 2-1 is a screenshot of Personal Space for presentation skills. In Personal Space, the student can watch the video and pause it at any time to write a comment and tag it by aspects. The submitted comment is displayed with the timestamp where the video was paused for writing that comment. An early study on AVW-Space (Mitrovic, Dimitrova, et al., 2017) provided evidence that commenting and using aspects prompted deeper thinking and self-reflection and helped the learners focus on the video content. Consequently, learners who commented on videos learned more than Passive students who only watched videos (Mitrovic, Dimitrova, et al., 2017). Therefore, nudges and interactive visualisations were proposed to foster commenting behaviour in AVW-Space (Mitrovic et al., 2019).

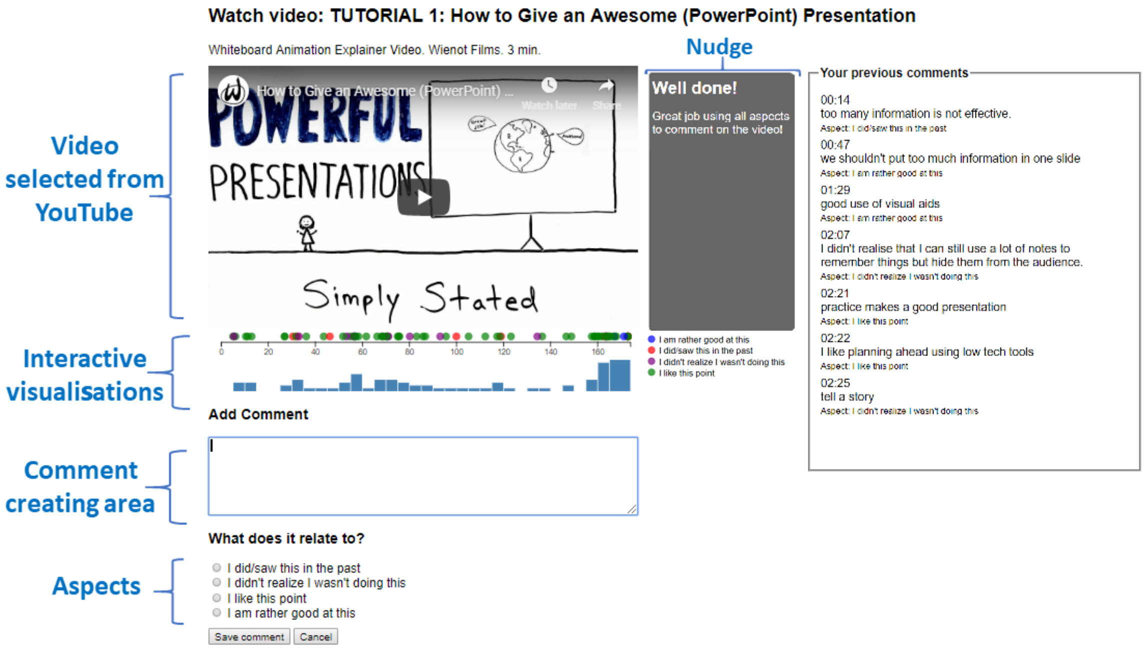

Figure 2-1 *Personal Space in AVW-Space*

Once the teacher approves comments to be shared anonymously among the learners, students can read and rate comments using the options defined by the teacher in the Social Space (Figure 2-2). Students can see but not rate their own comments. The rating activity improves learning through sharing learners' understanding and diverse points of view. An early study on AVW-Space revealed an increase in conceptual knowledge only for students who made and rated comments (Mitrovic, Dimitrova, et al., 2017). However, as seen in Figure 2-2, the learner confronts many comments to review and rate, many of which could be identical. For instance, some comments such as affirmations (e.g., "helpful") lack educational value for other learners to review. Therefore, reviewing an overwhelming number of low-quality, poorly structured comments could be frustrating. Hence, restructuring the list of comments based on their quality was suggested to direct students' attention to high-quality comments and enhance the usability of the rating task. Although initial supports (nudges and visualisations) were investigated in Personal Space to increase engagement, no support has been designed for the Social Space.

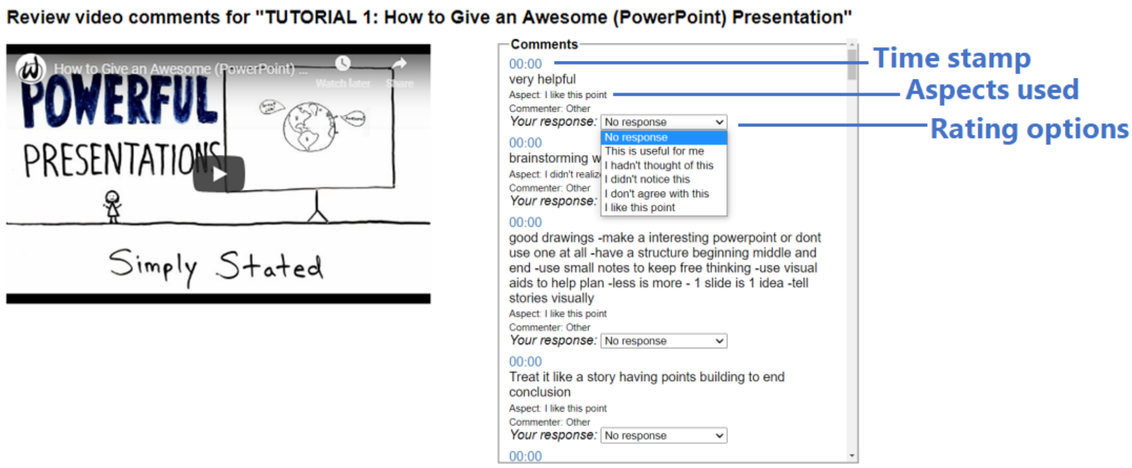

Figure 2-2 Social Space in AVW-Space

In all previous studies on using AVW-Space for training presentation skills, Survey 1 was administered at the beginning of each study. Survey 1 contained demographic questions, questions about the participant's knowledge of giving presentations, experience and training in giving presentations, and the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich & de Groot, 1990). The questions about training and experience in giving oral presentations, how often they watch videos on YouTube and how often they use YouTube for learning were based on the Likert scale from 1 (low) to 5 (high). However, MSLQ uses a Likert scale from 1 (lowest) to 7 (highest). Participants were instructed to watch and comment on the tutorial videos first, critique the example videos, and finally rate others' comments. At the end of each study, participants completed Survey 2, which included the same questions about giving presentations again to investigate whether students have increased their knowledge from Survey 1. Survey 2 also included two other questionnaires: NASA-TLX (Hart, 2006) to analyse the cognitive load in interaction with AVW-Space, and the Technology Acceptance Model (TAM) (Davis, 1989) to evaluate the perceived usefulness of AVW-Space. TLX-NASA scores are in the Likert Scale from 1 (lowest) to 20 (highest), but TAM questionnaire uses the Likert scale from 1 (highest) to 7 (lowest). In Surveys 1 and 2, the participants had 1 minute to list all concepts about the structure, visual aids, delivery, and speech. The students' answers were marked automatically, using the ontology of presentation skills developed in previous studies (Dimitrova et al., 2017). The marks for conceptual knowledge questions were used as pre- and post-tests scores (CK1 and CK2). It is noteworthy that the conceptual understanding of presentation skills is not the same as the target skill, but it was not feasible to organise full presentations for all participants in studies conducted in AVW-Space. Therefore, CK1 and CK2 are used in AVW-Space studies to measure and assess conceptual knowledge and learning.

Research on AVW-Space has operationalised the ICAP framework (Chi & Wylie, 2014) to investigate how the provided supports influence engagement in video-based learning. ICAP framework is based on students' observable actions as the primary indicator of their engagement. ICAP framework classifies students into four engagement levels: Interactive, Constructive, Active and Passive. Passive learners merely receive information without performing additional actions; they might listen to a lecture, read a book, or watch a video with no further visible engagement. However, Active learners perform additional actions, such as note-taking, but those actions simply repeat the received information (e.g., writing down lecturer's statements). On the other hand, Constructive learners generate additional information (e.g., a concept map or a self-explanation) that was not explicitly presented. Interactive learners contribute in discussions and collaborate with their peers and/or tutors to compare ideas and solve problems jointly. Chi & Wylie, 2014 provided evidence that the more engaged the students become, the more effectively their learning increases, i.e., Passive < Active < Constructive < Interactive. The Interactive category of ICAP is not applicable in AVW-Space, since it does not support collaboration. The previous investigation of comments (Hecking et al., 2017; Taskin et al., 2019) in AVW-Space showed that comments usually contain a summary of key points of videos, self-reflection and self-explanation. Hence, the operationalisation of ICAP for AVW-Space includes the following types of learners:

- **Passive students** log on to AVW-Space and watch videos but do not comment.
- **Active students** log on to AVW-Space, watch videos, and make comments that relate to the points in the videos.
- **Constructive learners** log on to AVW-Space, watch videos and write comments which add something new (e.g., reflection on personal experience or critical thinking about video key points).

Nudges and visualisations have been proposed to boost engagement levels regarding ICAP categories. The following subsections present more details about these two types of support and their limitations.

2.2.1 Nudges

AVW-Space provides nudges which are personalised prompts to foster commenting behaviour and to promote using diverse aspects while at the same time allowing learners to interact with videos freely and in their preferred manner. Nudges were initially introduced by (Thaler & Sunstein, 2008) as an unforced intervention to guide individuals to make better decisions while still giving them choices. AVW-Space follows the choice architecture framework (Münscher et al., 2016) and COM-B (Capability, Opportunity and Motivation - Behaviour) system (Michie et al., 2011) for designing the nudges (Dimitrova et al., 2017). Choice architecture defines techniques for selecting and presenting choices that lead to better behaviour. There are three main approaches in choice architecture:

- **Decision information:** focuses on representing information relevant to desirable decisions by rearranging information, simplifying them or providing a social reference point. In AVW-Space, some good examples of comments from previous students could be presented in the nudges.
- **Decision structure:** changes the structure of decisions by altering their formats. Changing default choices, grouping choices or changing choices consequences (e.g., benefit and cost) are examples of these techniques. In AVW-Space, a nudge could indicate a particular aspect that has been unutilised.
- **Decision assistance:** fosters self-regulation in decision-makers through reminders or supporting self-commitment. For AVW-Space, these nudges could be in the form of reminders for making a comment or positive feedback for using all aspects.

COM-B system defines three factors for behaviour change: 1) Capability to engage or having the essential knowledge and skills, 2) Opportunity of making the behaviour and extrinsic prompts, and 3) Motivation for getting directed to the behaviour. In the light of these factors, BCW (Behaviour Change Wheel) framework classifies different behaviour interventions into nine categories such as education, persuasion (stimulating an action), incentivisation or coercion and modelling (giving examples). AVW-Space adopted COM-B factors in the following way (Dimitrova & Mitrovic, 2021):

- **Capability**: Considering both the learner's self-regulation capabilities and their knowledge or experience in the target skill;

- **Opportunity**: Automatically identifying opportunities to support engagement in AVW-Space;
- **Motivation**: aiming to increase the learner's motivation to engage in AVW-Space and improve their knowledge.

AVW-Space devise nudges that increase engagement and trigger self-regulation and self-reflection using the adjusted COM-B system and choice architecture (Dimitrova et al., 2017; Mitrovic et al., 2019). These nudges are formulated as dialogue game $N = \langle G, T, I, O \rangle$ (Dimitrova & Brna, 2016; Dimitrova & Mitrovic, 2021), where *G* defines the pedagogical goal of the nudge; *T* represents the conditions where the nudge should be triggered; *I* means the interaction format (the message template) of the nudge; and *O* is the expected outcome behaviour after giving the nudge. The T element in the game dialogue of the nudges is defined based on the user's interactions data (e.g., timestamp of video being watched, comments made, and aspects used) (Dimitrova et al., 2017; Mitrovic et al., 2019). In other words, the nudges are personalised and provided to learners adaptively, based on their video-watching and commenting behaviour. Four Reminder nudges were initially added to AVW-Space, to encourage learners to write comments and use various aspects:

- **No Comment Reminder**: This nudge is initially triggered when a student passively watches a video to advise that writing comments is beneficial for learning.
- **No Comment Reference Point**: This is the next nudge the learner receives if they ignore the *No Comment Reminder Nudge*. This nudge will remind writing a comment and provide an example comment made by previous students to inspire the student.
- **Aspect Under-utilised**: This nudge is triggered when the learner has not used a specific aspect in commenting to suggest the learner writing comments using that aspect.
- **Diverse Aspects:** This nudge is positive feedback, triggered when a student has used all aspects in comments on the video (as presented in Figure 2-1).

An experiment on the effect of adding these nudges to AVW-Space revealed that Reminder nudges resulted in a higher number of comments, better usage of aspects, lower cognitive load in commenting and consequently, an increase in learning (Mitrovic et al., 2019). However, these nudges only focus on making comments and do not consider the content of the comments. Therefore, several analyses have been applied to the content of comments in AVW-Space to investigate how students perform differently in terms of comment content.

In an early study, the domain ontology of presentation skills was created (Dimitrova et al., 2017; Mitrovic, Dimitrova, et al., 2017). Then, the domain-specific concept ratio was used to infer how much a comment is relevant to the domain (Dimitrova et al., 2017; Mitrovic, Dimitrova, et al., 2017). The domain-specific ratio is the number of words from the domain ontology appearing in the comment, divided by the total number of words. However, in a case that a one-word comment has a word in the ontology, the value of domain-specific ratio will be 1; meaning the comment is highly relevant to the domain, while it does not necessarily convey any high-level thinking. Therefore, linguistic and psychological features of the comment were extracted from LIWC software (Pennebaker et al., 2015; Tausczik & Pennebaker, 2010) and were combined with the domain-specific ratio to predict whether the comment has the social value of comments. A high social value comment is one that receives the number of ratings in the top quartile of data. However, the quality of a comment is not necessarily attributed to its number of ratings since previous studies showed that comments displayed at the top of the reviewing list are more likely to be rated than the rest of the comments (Mitrovic, Dimitrova, et al., 2017).

A study on the vocabularies used in AVW-Space comments (Hecking et al., 2017) visualised the attention shift of comments content. This visualisation indicated that the concepts included in self-regulated learners' comments were aligned with the topics of the videos. However, students with lower self-regulatory skills made more affirmative comments and used fewer domain concepts in their comments. Additionally, this study applied network-text analysis to uncover the contextual pattern of comments made by various types of learners (Hecking et al., 2017). Network-text analysis (NTA) is a method for modelling the text as a network of concepts and extracting the relations between ontologies (Popping, 2000). The network-text analysis revealed that different types of learners could be characterised according to the number of domain-related terms they include in their comments. For instance, students with strong self-regulatory skills made several comments containing the majority of the domain vocabulary. In contrast, students with low self-regulatory skills wrote few comments using a small subset of the domain vocabulary. Moreover, terms used in self-regulated learners' comments exhibited a high degree of overlap. As a result, this study suggested presenting learners with the domain vocabulary to direct their attention to important concepts of the video and to facilitate finding appropriate words when commenting. Hence, a new research direction in AVW-Space is on topic modelling of video segments to automatically inform the learners about the concepts covered in each part of the video (Mohammed & Dimitrova, 2020). However, these analyses do not provide a model for the quality assessment of comments.

Recently, a study investigated the temporal and semantic alignment of comments with videos in AVW-Space to classify comments (Taskin et al., 2019). This study proposed three categories of comments: Simple (comments repeating a single point stated in the video), Summary (comments summarising points discussed in the video) and Elaborate (comments elaborating and reflecting on multiple points in the video rather than simply repeating). The majority of comments made by students were in the Simple category. In contrast, the number of Elaborate comments was ¼ of Simple comments. However, the pre- and post-test scores analysis revealed that students who made Elaborate comments learned significantly more than students who only made Simple comments. In addition, linear regression showed that the number of Reminder nudges students received is a significantly important predictor of the number of Simple comments students make. However, no significant relationship was found between the number of received Reminder nudges and Elaborate or Summary comments. In other words, the Reminder nudges increased the number of comments but did not improve the quality of comments.

The insights on various levels of elaboration in the content of comments inspired conducting an extensive analysis of comments quality and the development of nudges that encourage students to write better quality comments showing deeper reflection and elaboration. Offering nudges for writing better comments could improve the overall quality of comments, make the reviewing and rating task more beneficial (Mitrovic, Dimitrova, et al., 2017) and consequently increase engagement and learning.

2.2.2 Visualisations

AVW-Space presents visual learning analytics to encourage commenting. The visualisations are based on comments written by previous classes and show the distribution of comments throughout the video (Figure 2-1) (Mitrovic et al., 2019). The visualisations include the comment timeline and comment histogram. The comment timeline displays comments as coloured dots on the video timeline where the comments were created. The colour of a dot corresponds to the aspect used. The learner can see the comment text when hovering over a dot. In addition, clicking on the dot plays the video from the timestamp that the comment was made. Whenever a student receives an *Aspect Under-utilised Nudge* encouraging the student to use a particular aspect, the comment timeline visualisation changes to only show comments tagged by that particular aspect. However, the comments displayed in the visualisations are pre-defined and selected manually from previous studies with AVW-Space. The comment timeline visualisation provides social reference points to inspire ideas for commenting. The comment histogram demonstrates the number of comments made for the video sections to assist students in identifying important parts of the video, which received many comments.

The visualisations in AVW-Space support social learning by allowing students to see what other students wrote about the same video. However, a student may use the comment timeline just to learn others' perspectives and still make no comment. Thus, the timeline visualisation needs improvement to clarify its purpose: fostering commenting behaviour. Additionally, these visualisations do not provide information about the learner's progress to help them regulate their learning activities. Therefore, more provoking visualisations of the learner's model are required for boosting engagement in AVW-Space.

An early study showed that rating brings an additional benefit to learning on top of commenting (Mitrovic, Dimitrova, et al., 2017). Hence, visualising the student's progress in commenting and rating was suggested to motivate them to engage more actively. In addition, indicating the quality of students' comments was proposed for aiding students awareness of their engagement in commenting (Mitrovic, Dimitrova, et al., 2017). This study also revealed that students used opinion-inducing rating options (e.g., "I like this point" and "I do not agree with this") significantly more than the ones that trigger reflection and indicate learning new things (e.g., "This is useful for me", "I hadn't thought of this" and "I didn't notice this"). This finding highlighted the requirement for support which encourages using diverse rating options. Visualising received ratings could help students better understand the importance of rating constructively and using diverse rating options. Visualizing received ratings could also motivate the students to write high-quality comments by raising the students' awareness of their comments' influence on peers.

2.3 Quantitative Ethnography (QE)

The analyses of log data are crucial in learning analytics research since logs capture behaviour patterns over time (Paquette et al., 2021). There are methods for extracting temporal learning strategies from log data, such as clustering (Gasevic et al., 2017), sequence mining (Zhou & Bhat, 2021) and process mining (Shabaninejad et al., 2020). However, sequence mining only highlights local patterns, and process mining does not allow statistical testing. Quantitative ethnography (QE) is an effective approach for extracting patterns from log data which provides visualisations for qualitative interpretation along with statistical tests (Shaffer et al., 2016). QE (Shaffer, 2017) combines data mining, discourse analysis, statistics and ethnographic approaches to study the breadth of human behaviour (Wu et al., 2019). One common tool used in QE is Epistemic Network Analysis (ENA).

ENA is a network analysis technique that quantifies and models the co-occurrence of the coded data segments. The co-occurrences are modelled as weighted node-link networks. The generated networks allow visual and statistical comparisons between groups or samples (Shaffer & Ruis, 2017). The ENA algorithm constructs a network model for each line in the data using a moving window, showing how codes in the current line are connected to codes in the recent temporal context (Siebert-Evenstone et al., 2017). The temporal context is defined as the lines preceding the current line within a given conversation. The resulting networks are aggregated across all lines for each unit of analysis in the model. Networks are visualised as graphs where nodes correspond to the codes, and the weighted edges represent the relative frequency of co-occurrence, or connection, between two codes. A high co-registration correlation between the network and the projected space indicates that the positions of the network nodes can be used to explain the positions of plotted points in the space. Therefore, ENA allows comparing units in terms of their plotted point positions, individual networks, mean (centre) positions and mean networks, which average the connection weights across individual networks. Additionally, network difference graphs can be used to compare networks. These graphs are created by subtracting the weight of each connection in one network from the corresponding connections in another.

ENA is derived from the operationalisation of epistemic frame theory, which models learning as methods of thinking, acting, and being in the world of some community of practice (Shaffer & Ruis, 2017). Gamage et al. (2020) used ENA to compare novice MOOC learners with learners who had previous experience with MOOCs. ENA was applied to identify behavioural differences during video watching and social tasks, such as posting in discussion forums. Saint et al. (2020) combined ENA and process mining to provide richer insights on the self-regulated learning behaviours of low/high performing students in a learning management system. ENA was also used in an educational game for teaching complex systems in science to investigate how learners responded to the game events (e.g., feedback and new phenomena) and how their strategies evolved (Scianna et al., 2021). Karumbaiah et al. (2019) applied ENA on event logs from Physics Playground, an educational game for teaching Physics, to discover why some students quit the game.

Several studies have proved the effectiveness of AVW-Space in increasing engagement during VBL, using quantitative analyses of interaction data. However, there has been no ethnographic research on the longitudinal data of student interactions. Thus, this research investigates learning behaviours in AVW-Space and the effectiveness of supports in enhancing engagement more deeply by applying quantitative ethnography.

2.4 Conclusions

The most crucial challenge in video-based learning is maintaining and developing engagement due to the lack of direct human interaction, interactivity and personalised feedback. Several approaches have been suggested to tackle these challenges and enhance engagement, such as visualisations and personalised prompts. However, research on the effectiveness of these approaches in video-based learning is still in the early stage. AVW-Space is an online VBL platform that supports engagement via commenting and rating peers' comments. Although there have been initial prompts and visualisations to increase the number of comments in AVW-Space, there has been no support for boosting the quality of engagement in commenting and rating tasks. Also, previous research on AVW-Space investigated engagement only by quantitative analysis of activities and lacked ethnographic analysis of students' interactions. This PhD research addresses these gaps.

3 Quality Nudges Design

As discussed in the previous chapters, the first objective of this PhD research is to develop an automatic way to assess the quality of comments and design personalised nudges which encourage students to write better quality comments. The first step towards this goal is to develop quality schemes for comments and Machine Learning (ML) classifiers that can assess the quality of comments online, as the students write them. Then, the students' performance should be investigated by assessing the quality of their comments to design personalised Quality nudges which encourage students to write better quality comments. The Quality nudges need to be implemented in AVW-Space, in addition to the existing Reminder nudges. Therefore, this chapter addresses the following research questions:

RQ1. How can I define a robust scheme for human coders to assess the quality of comments?

RQ2. How reliable are the machine learning models trained using the quality scheme to classify comments based on their quality?

RQ3. How can I design and develop Quality nudges to encourage students to improve the quality of their comments?

After describing the datasets used for the assessment of comments and designing Quality nudges, I present two quality schemes for the comments on tutorial and example videos in Section 3.1. Next, Section 3.2 discusses the automation of quality assessment of comments using ML classifiers based on the quality schemes. After describing the design process of Quality nudges in Section 3.3, I present the implementation of the Quality nudges in Section 3.4. Finally, Section 3.5 summarises this chapter and discusses the challenges, limitations and future work for Quality nudges.

The data used for investigating comments quality and nudge design was collected from the previous AVW-Space studies conducted in a first-year engineering course (ENGR101) at the University of Canterbury in 2017, 2018 and 2019. This course used AVW-Space as an online resource for training students on presentation skills. In addition, the data from a study conducted with postgraduate (PG) students (Mitrovic et al., 2016) was used to evaluate the performance of the ML classifiers. All four studies used the same videos and aspects. The 2018 and 2019 studies shared the identical experiment design, with the control and experimental groups. In addition to the standard AVW-Space features (aspects, videos and rating categories), the experimental group in 2018 and 2019 studies received Reminder nudges (Mitrovic et al., 2019). However, the 2017 study (Mitrovic, Gostomski, et al., 2017) and the PG study did not include nudges.

3.1 RQ1: Quality Schemes

To assess the quality of comments, it is necessary to define different categories of comments. There have been several frameworks proposed for students' reflective writing. A study on academic reflective essays identifies different types of reflection such as personal belief, lessons learned or future intentions (Ullmann, 2017, 2019). Another framework for assessing the depth of students' self-reflections on their performance groups the contents into observation, effect, motivation and goal (Joksimović et al., 2019). A simulation environment for cross-cultural communications classifies the user's textual interactions with the system into different groups, such as statements on the situation and real-world stories (Dimitrova et al., 2013). However, a new labelling scheme for comments is needed for AVW-Space due to its special nature.

In order to develop quality schemes and ML classifiers for assessing the comment quality, comments made in previous studies were investigated. Table 3-1 presents the number of students who wrote comments and the number of comments on tutorial and example videos for each study. I designed two quality schemes, one for comments on tutorial videos and another for those on example videos.

Table 3-1 Number of students and comments in the previous AVW-Space studies

	PG	2017	2018	2019
Participants	32	158	192	189
Comments on tutorial videos	346	670	1,144	1,101
Comments on example videos	368	575	687	660

The proposed quality schemes for comments on tutorial and example videos include ordinal categories, meaning the first category has the lowest quality, and the last category shows the highest quality. The ordinal categories are also used in the ICAP framework and other research focusing on higher-order thinking (X. Wang et al., 2016). Table 3-2 presents the categories of comments on tutorial videos, with some examples from previous AVW-Space studies. This scheme contains five categories: (1) Affirmative, negative or off-topic, (2) Repeating, (3) Critical and analytical, (4) Self-reflective and (5) Self-regulating comments. Comments in categories 1 and 2 are pedagogically undesirable since they do not convey deep thinking about the videos. However, comments in category 3 show more critical thinking about the video, as learners elaborate on the video content. In category 4, learners reflect on their previous experience in relation to the video. Finally, students indicate a high level of learning in category 5 by planning how to improve their future presentations using the ideas covered in the videos. In this thesis, I refer to comments in categories 3, 4 and 5 as high-quality comments and consider those in categories 1 or 2 as low-quality ones.

Table 3-2 Quality scheme for comments on the tutorial videos

Category	Definition
1. Affirmative, negative, off-topic	Comments which are irrelevant or merely affirmative or negative with no explanation. e.g., [Aspect: I did/saw this in the past] *"very helpful."*
2. Repeating	Comments which only repeat the video content. e.g., [Aspect: I like this point] "*limit each slide to one key idea.*"
3. Critical and analytical	Comments which mention points that are implicitly covered in the video, or show critical thinking on the content of the video. e.g., [Aspect: I like this point] "Presentations can be boring and long whereas stories are more enjoyable and can have clear direction if formulated properly."
4. Self-reflective	Comments in which the learner reflects on their behaviour and previous experience or knowledge on giving presentations. e.g., [Aspect: I saw/did this in the past] "My past speeches have had very interesting beginnings."
5. Self-regulating	Comments where the learner decides what they would do to improve themselves in future. e.g., [Aspect: I didn't realize I wasn't doing this] "I will definitely be trying to smile more throughout my next presentation."

I also designed a quality scheme for comments on example videos (Table 3-3), which is similar to the scheme for tutorial videos with the exception of self-regulating and self-reflective categories. The reason for excluding those two categories is that the students were instructed to critique example videos using aspects (structure, visual aids, delivery and speech).

Table 3-3 Quality scheme for comments on the example videos

Category	Definition
1. Affirmative, negative, off-topic	Comments which are irrelevant or merely affirmative or negative with no explanation. e.g., [Aspect: Visual aids] "*This was helpful.*"
2. Repeating	Comments that list or name good/bad practices in the presentations without explaining the effects and causes of the practice. e.g., [Aspect: Speech] "*End on a question*."
3. Critical and analytical	Comments which criticise examples, explain the effect of a good/bad practice in the presentation or offer advice for improvement. e.g., [Aspect: Speech] "Should give more meaning to the statistic by placing it in context."

To evaluate the proposed quality schemes, 167 comments from 2018 and 2019 studies (110 and 57 comments on tutorial and example videos, respectively) were selected via stratified sampling since these two studies had an identical experimental design. Then, three expert coders labelled those comments independently. Next, ordinal Krippendorff's α ("Krippendorff's Alpha," 2010) was used to evaluate the inter-coder agreement. Krippendorff's α values were .78 and .69 for tutorial and example videos, respectively. These Krippendorff's α values are higher than .66, the suggested minimum acceptable value of α for inter-coder agreement ("Krippendorff's Alpha," 2010). Therefore, RQ1 was addressed. Then, the three coders reviewed the comments on which they disagreed, and the definitions in the schemes were clarified for further manual classification. Finally, I labelled all comments from previous studies using the quality schemes.

3.2 RQ2: Automating the Assessment of Comment Quality

This section first presents the extraction of various linguistic and psychological features from comments and investigates whether these features are useful for identifying each quality category. Next, various machine learning classifiers are explored to automate the quality assessment of comments. Finally, the generalisability of the best performing quality assessment models is discussed by evaluating their performance on unseen datasets.

3.2.1 Feature Analysis of Comments

The analysis of written self-reports can provide insights into students' conceptual comprehension and engagement. Earlier research in this area focused on the textual analysis of students' essays (Crossley et al., 2019; McNamara et al., 2013), which are longer and more structured than video annotations. Text analysis has also been applied to students' answers to questions asked during or after teaching sessions (Arbogast & Montfort, 2016; Heilman & Madnani, 2015; Prevost et al., 2013). However, the answers to specific questions are potentially more restricted than students' video annotations, because annotation can cover concepts from any part of the video. In addition, there have been several studies on textual analysis of online forum discussions due to the increasing use of MOOCs (Crossley et al., 2015; Martín-Monje et al., 2018). Although forum discussions are comparable to video annotations in terms of length and open-endedness, video annotations are less conversational than forum posts. These differences between comments and other types of written self-reports highlight the need for further dedicated analysis of video annotations.

Text categorisation is a common text analysis technique in education, which is widely used in automatic scoring (Ferreira-Mello et al., 2019). Python provides the most suitable packages for educational text analysis (Cunningham-Nelson et al., 2018). Another technology focusing on text parts of speech, sentence structure, and semantic word category is Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010). LIWC is utilised mainly for counting the number of words in various linguistic and psychological categories. The tools used for feature analysis and machine learning approaches in this research are LIWC, SPSS[2], Weka[3] and Scikit-learn[4] library in Python.

The comments first needed to be converted into numerical features. The features were extracted from the comments of the 2018 and 2019 studies, since their experimental design was identical. Overall, there were 1,347/2,245 comments on example/tutorial videos. The word-count approach was taken (Tausczik & Pennebaker, 2010) for extracting numerical features rather than full parsing, since these comments are not always grammatically correct. Therefore, the comment text is converted to numbers by counting the frequencies of words (or their stems) in the LIWC dictionaries. LIWC dictionaries are collected from various psychological constructs such as cognitive processes. Moreover, LIWC can calculate linguistic elements, including pronouns, verbs and their tenses. The previous studies on reflective writing in various contexts also used LIWC features (Joksimović et al., 2019; M. Liu et al., 2019). In addition to the LIWC features, I used the domain-specific ratio and unique domain-specific ratios proposed in a previous study (Mitrovic et al., 2019) to consider the topic of comments. The unique

[2] https://www.ibm.com/marketplace/spss-statistics
[3] https://www.cs.waikato.ac.nz/ml/weka/
[4] https://scikit-learn.org/

domain-specific ratio is the number of unique words from the domain vocabulary appearing in the comment, divided by the total number of words in the comment. For comments on tutorial videos, a binary feature was also derived for indicating whether the used aspect is reflective. As can be seen in Figure 3-1, comments in categories 4 and 5 (self-reflective and self-regulating comments) were mostly tagged by reflective aspects ("I didn't realise I was doing this", "I am rather good at this" or "I did/saw this in the past"), while comments in lower categories were mostly tagged by the "I like this point" aspect which is non-reflective. Therefore, it is necessary to consider the type of used aspect to assess the quality of comments on tutorial videos.

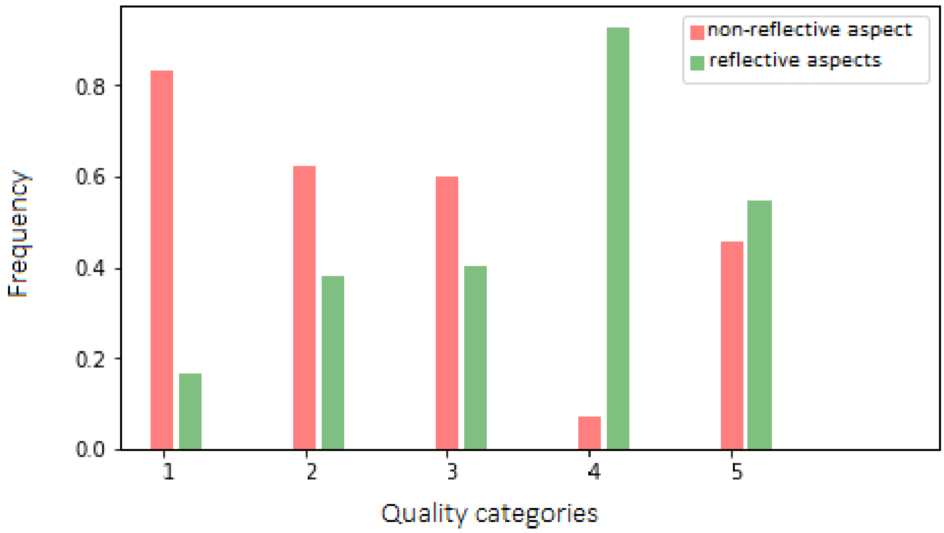

Figure 3-1 Frequency of aspects tagged for comments on tutorial videos in different quality categories

To reduce the size of the feature set, I removed LIWC features that are not meaningful in this context, such as "biological" words (e.g., eat, blood, pain) or "personal concerns" (e.g., work, religion, death). I also removed LIWC summary features since they are derived from primary features (see Appendix A – LIWC Features). Therefore, the selected LIWC features are:

word count, word per sentence, six-letter (or more) words, dictionary words, function words, pronouns, personal pronouns (e.g., I, we, you, she/he, they), impersonal pronouns, article, prepositions, auxiliary verbs, adverbs, conjunctions, negations, verbs, adjectives, comparisons, interrogatives, numbers, quantifiers, affective processes, positive emotions, negative emotions, social processes, cognitive processes, insight, causation, discrepancy, tentative, certainty, differentiation, perceptual processes (e.g., see, hear, feel), drives, affiliation, achievement, power, reward, risk, focus on past, focus on the present, focus on future, relativity, informal language, swear words, net-speak, assent, non-fluencies and fillers.

After normalising features to values between 0 and 1 using Min-Max scaling, I investigated the correlation of the LIWC features with the quality of comments to better understand the data. The correlation analysis for tutorial comments showed that features like "I" pronoun (r = .34), number of words (r = .22) and number of words per sentence (r = .23) have the highest positive correlations with the quality category of comments. However, there was a high negative correlation between positive emotions and the category of comments (r = -.21). This could indicate the short affirmative comments in category 1 (e.g., "This was helpful."). In comments on example videos, there are relatively high correlations between the quality of the comments and the total number of words (r = .45) and the number of words per sentence (r = .49). However, there was a high negative correlation between positive emotions and the category of comments made on example videos (r = -.31), similar to the comments made on tutorial videos.

I investigated whether the features capture the same quality categories agreed by the expert coders. First, the features centroids for each category were computed. Then, K-means algorithm was utilised to cluster the comments on tutorial and example videos using LIWC features and unique domain concepts ratio. I examined clustering with different numbers of clusters to 1) find an appropriate cluster count which can differentiate comments, and 2) investigate whether these generated clusters overlap with categories in the quality schemes. Thus, the silhouette score for each cluster count was computed to measure how well-partitioned they are. Silhouette analysis compares the intra-cluster distance of a sample with the mean nearest-cluster distance (Rousseeuw, 1987).

Figure 3-2 represents the change of silhouette scores by increasing the number of clusters. I chose 5 clusters for tutorial comments and 3 for example comments because the higher number of clusters did not show any significant increase in the silhouette score. These selected cluster counts are equal to the number of categories in the proposed schemes. Since the silhouette scores of 3 and 4 clusters are higher than that of 5 clusters in tutorial comments, it is worth analysing further whether these clusters are appropriate for describing the quality of tutorial comments.

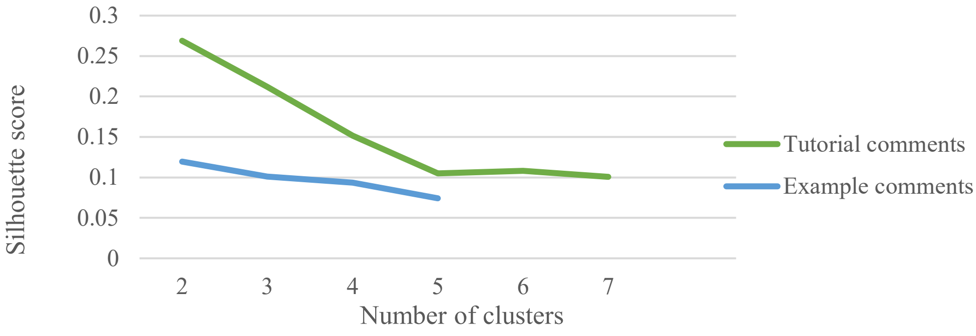

Figure 3-2 Silhouette score for different number of clusters of comments

After clustering comments into the chosen number of clusters, I compared the centroid of each cluster with the features centroids in each category. The cluster centroids overlapped approximately with feature centroids of the categories and described the same behaviours as the proposed categories. In other words, the chosen features can spot the patterns introduced in the schemes.

3.2.2 Training ML Classifiers

In order to train and evaluate ML classifiers for predicting comments quality, the comments from the 2018 and 2019 studies were split into the training (80%) and test (20%) sets on the student level. That is, the students whose comments are in the training set have no comment in the test set. However, I noticed that the data is imbalanced, as the numbers of comments in each category are different, as illustrated in Table 3-4 for the training set. For example, among the tutorial comments, there are 924 comments in category 2, but there are only 41 comments in category 5.

Table 3-4 Categories distribution in tutorial and example comments in the training set

	Tutorial comments					Example Comments		
Category	1	2	3	4	5	1	2	3
Comments	48	924	395	387	41	27	668	320

One of the techniques to tackle the challenge of imbalanced data could be random down-sampling the categories with higher frequency and random up-sampling categories with a lower frequency. However, due to the small number of comments in categories 4 and 5, up-sampling would duplicate samples of these categories in training and test sets of cross-validation folds. Therefore, sampling is not suitable for this type of data set. Instead, I set the class weights reversely proportional to the frequency of classes. Thus, categories 1, 4 and 5 have higher weights than categories 2 and 3 in tutorial comments. The emphasis on category 1 is sensible in this context, since a low-quality comment needs to be detected accurately to encourage its writer to improve comment writing. Besides, it is preferable to avoid providing interventions to students who have already written good comments in categories 4 or 5. This interpretation is also applicable for comments on example videos, where category 1 has the highest weight and requires the most emphasis in providing interventions. However, there are also many other techniques to deal with imbalanced data, such as anomaly (outlier) detection and cost-sensitive learning, which will be discussed later.

To automate the quality assessment of comments, I examined various traditional classifiers rather than new deep learning approaches due to insufficient data and the low interpretability of deep learning approaches. Some simple, commonly applied algorithms in text classification are Support Vector Machine (SVM) and tree-based algorithms such as decision trees (Kowsari et al., 2019).

First, a decision tree classifier was trained to discover the essential features in this dataset. The depth limit of eight was set to control the complexity of the models. The root of the generated tree for tutorial comments was the singular first-person pronoun (“I”). Comments with a higher value of this feature are likely to be in categories 4 or 5, in which students reflect on themselves. At the next level of this classifier tree, LIWC features identifying causation, emotions, discrepancy and unique domain-specific ratio were the indicators of the categories. For example comments, word per sentence was the root of the trained decision tree. This feature separates comments in category 1 from others where students write longer sentences to express their thoughts on the videos. In the next level of this tree, features such as causation, negation, emotions and perceptual processes were the nodes of the decision tree.

I also applied SVM with the linear kernel and one-versus-one schema for more interpretability. The trained classifiers had higher coefficients for personal pronouns ($\beta = 3.14$), affective processes ($\beta = 2.52$) and unique domain-specific ratio ($\beta = 2.28$) and discrepancy ($\beta = 1.81$) for tutorial comments. The highest coefficients of the SVM trained on example comments were for comparisons ($\beta = 2.06$), interrogatives ($\beta = 1.65$) and differentiation ($\beta = 1.50$). However, SVM and decision tree classifiers assume no meaningful order for the quality categories. Therefore, I examined ordinal classifiers.

I attempted the Threshold-based Ordinal Regression (Pedregosa, 2015; Rennie & Srebro, 2005), which considers k-1 thresholds to partition the predictor value range to k segments corresponding to k ordinal labels. These thresholds are trained by minimising the cost of threshold violation. There are two cost function constructions introduced for penalising threshold violation:

- **Immediate-threshold (IT)** which only minimises the threshold violation in two subsequent thresholds indicating the correct segment for each ordinal label.
- **All-threshold (AT)** which sums the cost of all threshold violations for each label. Thus, the higher distance between the predicted and actual labels causes a greater cost.

I also evaluated another ordinal classification approach which converts the problem of k ordinal classes ($V_{1,2,\ldots,}V_k$) into $k-1$ binary class problems corresponding to k-1 first ordinal classes. That is, the binary class problem corresponding to V_i (where $1 \leq i \leq$ k-1) predicts $Pr(A* > V_i)$, the probability of having the class attribute $A*$ higher than the corresponding category (V_i). This probability can be calculated using any probability predictor. After solving the k-1 binary class problems, the probability of class attribute equal to each ordinal class ($Pr\ (A* = V_i)$), can be derived from the probabilities computed in k-1 binary class problems ($Pr\ (A* > V_i)$). Having the probability of each k ordinal class, the class with the highest probability is selected as the class of the instance (Frank & Hall, 2001).

I evaluated these models by the weighted F1-score as recommended for assessing performance in imbalanced data (Jeni et al., 2013). In addition, I measured the Mean Squared Error (MSE) for the trained models due to the desire of finding the model with the lowest distance between predicted and the actual categories. Table 3-5 shows the mean of measurements in 10-fold cross-validation for applied algorithms on tutorial and example comments. The baseline performance using Zero Rule (Zero-R) is also provided to compare the performance of other algorithms. Zero Rule predicts all the instances to be in the category with the highest frequency (category 2 in this problem).

Table 3-5 Performance of different ML models using cross-validation

Model	Tutorial Comments				Example comments			
	MSE	Recall	Precision	F1	MSE	Recall	Precision	F1
Zero-R	.64	.66	.44	.53	.47	.48	.21	.35
Decision Tree	.99	.45	.58	.48	.64	.51	.54	.52
SVM	.99	.55	.68	.58	.61	.59	.64	.60
Immediate-threshold	1.35	.35	.59	.39	.53	.53	.56	.53
All-threshold	.91	.46	.60	.50	.50	.55	.57	.55
k-1 bi-classifiers (using SVM)	.50	.69	.66	.65	.41	.61	.64	.57

As seen in Table 3-5, identifying class weights based on the frequency of classes resulted in unsatisfactory performance. Therefore, I took a cost-sensitive approach to define proper costs for different misclassifications regarding educational purposes. For example, misclassifying a comment in category 1 as 3 is worse than misclassifying a comment in class 3 as 5, because comments in category 1 show very limited engagement and therefore must be correctly detected and supported appropriately. Thus, the cost of misclassifying a comment in category 1 as 3 should be higher. I designed the cost matrices presented in Figure 3-3, which still consider the order of categories by increasing costs as the misclassification distance grows. In these matrices, misclassifications for categories 1 and 5 for tutorial videos and categories 1 and 3 for example videos, have higher costs since the comments in category 1 need educational support the most. At the same time, a high-quality comment does not require pedagogical interventions. After many experiments with this approach, random-forest was selected as the well-performing base classifier. I refer to these classifiers as T_a and E_a, for comments on tutorial/example videos, respectively.

$$\begin{array}{r c} classified\ as: & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} actual = 1 \\ actual = 2 \\ actual = 3 \\ actual = 4 \\ actual = 5 \end{matrix} & \begin{pmatrix} 0 & 20 & 40 & 60 & 80 \\ 10 & 0 & 10 & 20 & 30 \\ 20 & 10 & 0 & 10 & 20 \\ 30 & 20 & 10 & 0 & 10 \\ 80 & 60 & 40 & 20 & 0 \end{pmatrix} \end{array} \qquad \begin{array}{r c} classified\ as: & \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{matrix} actual = 1 \\ actual = 2 \\ actual = 3 \end{matrix} & \begin{pmatrix} 0 & 25 & 50 \\ 10 & 0 & 10 \\ 50 & 25 & 0 \end{pmatrix} \end{array}$$

Figure 3-3 Selected cost matrices for comments on the tutorial (left) and example videos (right)

Table 3-6 reports the weighted mean of metrics for evaluating the performance of the cost-sensitive classifiers. F1-score is recommended for evaluating ML models on imbalanced data (Jeni, Cohn, & De La Torre, 2013). In addition, the average cost and cost-saving were calculated to evaluate models in terms of costs. Cost-saving (Equation 1 cost-saving definition) is the fraction by which the actual predictions reduce the costs in the worst case of misclassification. I started with models E_a and T_a which were trained to identify each category from the corresponding quality scheme. Classifier T_a identifies comments belonging to one of the five categories in the quality scheme for the tutorial videos. The performance of classifiers T_a and E_a was not satisfactory. Therefore, I considered whether it is necessary to be able to predict each category of comments individually. For example, categories 4 and 5 of the tutorial comments show a very high reflection level; in such cases, a pedagogical intervention is not needed. Positive feedback would be encouraging when the student initially writes a comment belonging to one of those two categories, but providing positive feedback on each high-quality comment would not be appropriate for well-performing students. For that reason, I considered various combinations of quality categories and trained different classifiers (Table 3-6). For comments on example videos, the only category requiring an immediate intervention is category 1. However, there is no need for distinct interventions when a student writes comments in category 2 or 3 on example videos, and positive feedback would be sufficient. Therefore, I grouped example comments from categories 2 and 3 together; the resulting classifier E_b differentiates between two types of comments (category 1 versus the union of categories 2 and 3). For tutorial comments, I explored various groupings resulting in classifiers T_b to T_d.

$$cost\_saving = 1 - \frac{cost\ of\ predictions}{Maximum\_Cost}$$

Equation 1 cost-saving definition

As can be seen in Table 3-6, classifiers T_c, T_d and T_e have better performance than the initial model (T_a). Classifier T_d aligns well with the ICAP framework (Chi & Wylie, 2014): active learning (just repeating the received information) is represented by categories 1 and 2; constructive learning (adding information that was not explicitly taught) is captured by categories 3, 4 and 5. However, having only 2 categories is not enough for capturing different behaviours and providing adequate support. Classifier T_c is similar to T_d, but it distinguishes between categories 1 and 2 to provide proper support. Classifier T_e groups comment into "off-topic/short affirmative or negative", "reflecting on the video", and "reflecting on personal experience and self-regulating". For comments on example videos, classifier E_b predicts whether a comment describes the video and analyses the strengths or weaknesses of the presentation in the video. E_b and T_e were selected as the best-performing classifiers, and their cost matrices are presented in Figure 3-4.

Table 3-6 Performance of the cost-sensitive models on the test set

Video	Model	TPR	FPR	Precision	F1-score	Avg. Cost	Cost-saving
Example	E_a: 1, 2, 3	.71	.22	.75	.71	3.79	.85
	E_b: 1, 2+3	.95	.21	.97	.96	.99	.98
Tutorial	T_a: 1, 2, 3, 4, 5	.72	.18	.72	.68	3.53	.877
	T_b: 1, 2, 3, 4+5	.70	.18	.72	.64	4.42	.873
	T_c: 1, 2, 3+4+5	.80	.17	.80	.80	2.86	.886
	T_d: 1+2, 3+4+5	.74	.26	.80	.73	3.10	.877
	T_e: 1, 2+3, 4+5	.84	.15	.86	.84	2.08	.881

$$\begin{array}{ll} classified\ as: & \begin{matrix} \;\;1 & \;\;2+3 & \;4+5 \end{matrix} \\ \begin{matrix} actual = 1 \\ actual = 2+3 \\ actual = 4+5 \end{matrix} & \begin{pmatrix} 0 & 20 & 40 \\ 10 & 0 & 10 \\ 40 & 20 & 0 \end{pmatrix} \end{array} \qquad \begin{array}{ll} classified\ as: & \begin{matrix} \;1 & \;2+3 \end{matrix} \\ \begin{matrix} \text{actual} = 1 \\ \text{actual} = 2+3 \end{matrix} & \begin{pmatrix} 0 & 50 \\ 10 & 0 \end{pmatrix} \end{array}$$

Figure 3-4 Cost matrices of T_e and E_b classifiers

Figure 3-5 presents the confusion matrices for the selected classifiers. Classifier T_e can identify comments in category 1 correctly. The only misclassification in category 1 is for a comment saying "hello Jim" (Jim is the name of a character in a tutorial video). Besides, some of the very short comments in higher quality categories were misclassified as 1. For instance, "guideposts" should be classified as 2+3, but it is misclassified as 1. The F1-scores for classes 1, 2+3 and 4+5 are .72, .89 and .72, respectfully. Classifier E_b classified most comments in category 1 correctly, but misclassified thirteen comments of class 2+3. The two misclassifications in category 1 are for comments that use domain-specific concepts to discuss the subject of the example rather than criticising the presentation skills of the presenters. For example, "valid points" is misclassified as 2+3 since the comment includes "valid" and "points", which are two concepts from the domain ontology.

$$\begin{array}{ll} \textit{Classified as} & \begin{matrix} \;1 & \;2+3 & \;4+5 \end{matrix} \\ \begin{matrix} actual = 1 \\ actual = 2+3 \\ actual = 4+5 \end{matrix} & \begin{pmatrix} 8 & 1 & 0 \\ 4 & 283 & 47 \\ 1 & 18 & 85 \end{pmatrix} \end{array} \qquad \begin{array}{ll} \textit{Classified as} & \begin{matrix} \;1 & \;2+3 \end{matrix} \\ \begin{matrix} actual = 1 \\ actual = 2+3 \end{matrix} & \begin{pmatrix} 7 & 2 \\ 13 & 310 \end{pmatrix} \end{array}$$

Figure 3-5 Confusion matrices for T_e (left) and E_b (right)

3.2.3 Generalisability of the Classifiers

I investigated the generalisability of the selected models (T_e and E_b) by evaluating their performance on unseen comments from the 2017 and PG studies. The 2017 study was done with a similar population of students (the 2017 class of the same course), who have not received nudges. Therefore, I expected the performance of the classifiers to be similar to that on the test set. On the other hand, PG students usually have much stronger learning and metacognitive skills. Therefore, I expected that the performance of the classifiers on the PG data would be worse than their performance on the training/test set.

The performance of selected classifiers on the data from 2017 and PG studies is reported in Table 3-7. The classifiers performed differently for PG and 2017 datasets because the distributions of the combined categories in these studies are different from those in the 2018 and 2019 studies, as highlighted in Table 3-8. Postgraduate students wrote more reflective comments (categories 4 and 5) than the other groups of students. Also, the PG students made slightly more high-quality comments on example videos (categories 2 and 3) than first-year students. When looking at the first-year students only, the 2017 set differs from the 2018 and 2019 datasets; the provision of nudges in the 2018 and 2019 studies resulted in more reflective comments (Mitrovic et al., 2019).

Table 3-7 Performance of merged categories models on 2017 and PG data

Data	Model	TPR	FPR	Precision	F1-score	Avg. Cost	Cost-saving
2017	E_b: 1, 2+3	.93	.26	.96	.94	1.54	.88
	T_e: 1, 2+3, 4+5	.72	.26	.81	.74	3.58	.77
PG	E_b: 1, 2+3	.96	.85	.96	. 96	2.08	.82
	T_e:1, 2+3, 4+5	.68	.30	.70	.69	4.61	.79

Table 3-8 Percentages of merged categories in different data sets

Video	Categories	Training	Test	2017	PG
Example	1	2.66 %	2.72 %	5.91 %	2.16 %
	2+3	97.33 %	97.28 %	94.09 %	97.83 %
Tutorial	1	2.67 %	2.01 %	3.58 %	.29 %
	2+3	73.48 %	74.72 %	79.40 %	58.95 %
	4+5	23.84 %	23.26 %	17.01 %	40.75 %

This section addressed RQ2 by evaluating the performance of quality assessment models for comments on tutorial and example videos (T_e and E_b). The assessment of comment quality is a starting point towards enhancing tailored support for engagement in AVW-Space. The quality assessment models enable us to design personalised nudges that focus on the quality of comments a student writes them. For instance, when a student submits a comment in category 1, AVW-Space would provide an immediate nudge to encourage the learner to be more focused on the video content in their comment. In addition, a nudge could suggest more elaboration or self-reflection when a student submits a comment in category 2 or 3. The system should also give positive feedback to the student who writes a self-reflective or self-regulating comment. The following section discusses how Quality nudges can be designed using the quality assessment models.

3.3 RQ3: Designing Quality Nudges

I followed the choice architecture-driven framework (Münscher et al., 2016) to design nudges that can lead to the desired behaviour of making high-quality comments while allowing the students to make free choices. As mentioned in Chapter 2, choice architecture includes three basic categories of intervention techniques corresponding to decision-making stages: decision information, decision structure, and decision assistance. Decision information technique can nudge students towards making self-reflective or critical thinking comments by explaining and giving them an example of such comments made by others. At the end of each video, the decision structure technique can present a default prompted choice, such as encouraging students to synthesise the main points and connect them to their own experiences. An example of the decision assistance technique is fostering deliberate commitment to writing high-quality comments by providing positive feedback when a student makes self-reflective or self-regulating comments frequently.

Each Quality nudge is represented in the form of a game dialogue: $N = <G, T, I, O>$ as described in Chapter 2 (Dimitrova et al., 2017; Dimitrova & Mitrovic, 2021). In a game dialogue, *G* is the pedagogical goal of the nudge, *T* represents the conditions when the nudge should be triggered to the student, *I* is the interaction format or the template message of the nudge, and *O* is the expected outcome in terms of the desired behaviour change after the nudge. The nudge conditions evaluate the student profile, the history of nudges the student received for the current video, the quality of comments the student made on the current video, and the time in the video that the student is watching. I designed seven Quality nudges for tutorial videos and three Quality nudges for example videos.

3.3.1 Quality Nudges for Tutorial Videos

To design Quality nudges for tutorial videos, I analysed the relationships between the students' performance on commenting and their profiles in the 2019 and 2018 studies. The student profile contains information about the level of training and experience in giving presentations and their MSLQ scores from Survey 1. The performance of each student was identified using the following boolean attributes, based on the quality assessment of their comments:

1. The student wrote at least one comment in category 1

2. More than 50% of the student's comments are reflective or self-regulating comments (i.e., category 4 or 5)
3. The student wrote a comment in category 2 or 3, but the comment is vague. Vague comments are short and contain words like "thing" or "that" without mentioning what they are referring to.
4. More than 50% of the student's comments repeat the video content (i.e., category 2)
5. More than 50% of the student's comments show critical thinking (i.e., category 3).

These attributes are not mutually exclusive, since a student can have a comment in category 1 and many comments in category 2 or 3, at the same time. These performance attributes help in designing the Quality nudges that students need to receive and their conditions. For instance, if a student has a comment in category 1, they should receive a nudge immediately to elaborate more. When a student writes reflective comments frequently, a nudge in the form of positive feedback should be given. However, when a student has a vague comment, a nudge should ask the student to clarify more. In addition, a student who writes repeating comments frequently should be asked to think more critically. Finally, a student who makes critical comments frequently should be nudged toward writing reflective comments.

To identify personalised Quality nudges which suit students learning strategies, I analysed the correlation between the performance attributes and the student profiles from the 2018 and 2019 studies. I found that training and MSLQ self-efficacy scores have a significant negative correlation ($r = -.19$, $p < .01$) with performance attribute 3 (i.e., repeating the video content in more than 50% of comments). Also, there was a significant correlation ($r = .13$, $p < .05$) between critical thinking and elaboration scores from MSLQ with performance attribute 5 (i.e., more than 50% comments show critical thinking). Therefore, these profile features (i.e., training and MSLQ scores for self-efficacy and elaboration) were selected for designing nudges.

To understand the linguistic differences of the vague, repeating and critical thinking comments, I analysed their LIWC features. Word count, unique domain-specific ratio, domain-specific ratio, insight, cause, assent and tentative features were selected as the features with the highest chi-squares in vague, repeating and critical thinking comments. Next, I trained a rule-based classifier, JRip (Cohen, 1995), on the training set (N = 1,317) to extract the association of the selected linguistic and student profile features with these groups of comments. The classifier used the selected LIWC features of comments, the predicted comment quality and the selected profile features of the student who made the comments. This rule-based classifier performed well on our test data (N = 260); (F1-score, recall and precision = .86). The rules derived from the classifier were used in the design of nudges to know whether a comment predicted as category 2+3 is vague, repeating or showing critical thinking. Some of the derived rules are based on the student's MSLQ scores, and some are purely based on the comment's LIWC features and domain-specific ratio. Therefore, a student who has not answered the MSQL questions still receives nudges based on the rules that only focus on the comment content.

Finally, I designed the Quality nudges for tutorial videos, focusing on guiding students to improve the quality of their comments toward self-reflection and self-regulation, which are the highest quality for tutorial videos. These Quality nudges are evaluated immediately after the student submits a comment. The Quality nudges for tutorial videos are as follows:

Table 3-9 Definitions of Quality nudges for tutorial videos

Name	Goal	Trigger Conditions	Interaction Template	Outcome
Elaborate Nudge	To prevent short or off-topic comments.	The submitted comment is predicted as category 1.	*"Try to elaborate more on the video in the next comment. For example:* [An example comment of category 3, manually selected from previous studies]*"*	The student writes longer and more relevant comments.
Vague Comment Nudge	To prevent ambiguity in comments	The comment is predicted to be in category 2 or 3, but it is short and has words like "thing" or "that" without mentioning what they are referring to. The vague comments are defined by the extracted rules from the JRip classifier as short comments: word count < 3 and very low LIWC scores for "insight" or "cause".	"This comment sounds a bit vague; can you clarify more in your next comments?"	Future comments are less ambiguous.
Repeating Comment Nudge	To reduce the number of comments that just repeat the video content, and encourage	The comment is predicted to be in category 2 (repeating the video content), and more than three out of the last five comments on the current video are in category 2 or 3. The second condition is to avoid giving the same nudge repetitively.	"Great! It seems you're very focused on the video. Can you think of the drawbacks or benefits of these tips?"	The student will think more critically about the points made in the video.

	critical thinking			
Critical Thinking Nudge	To encourage the student to engage in self-reflection and self-regulation.	The comment is predicted to be in category 3 (showing critical thinking), and the student has written comments in category 2 or 3 more than three times in the last five submitted comments on the current video.	"Great thinking! Now that you've learned this, can you reflect on your previous experience? What can you plan to improve your presentation skill? For example: [An example comment in category 4 or 5 (self-reflection or self-regulation), selected from previous studies]".	The student thinks about their previous experience or plans for improving future presentations.
First Self-reflection Nudge	Encourage the student to write more self-reflective comments.	The comment is predicted to be in category 4 or 5 (self-reflection or self-regulating), and it is the first comment in these categories for the current video.	"Well done, you made your first self-reflective comment for this video! Self-reflections help in improving your soft skills."	The student writes more self-reflections.
Frequent Self-reflection Nudge	Encourage the student to keep writing self-reflective comments.	The comment is predicted to be in category 4 or 5 (self-reflection or self-regulating), and the student has written more than three comments in category 4 or 5 in the last five comments.	"Awesome! You have made a lot of self-reflective comments! Keep writing them."	The student develops the habit of writing more self-reflections.
Final Self-reflection Nudge	Encourage the student to write self-reflections after watching the video.	The student is watching the last 30 seconds of the video, and has not written any comment in category 4 or 5 (self-reflection and self-regulating) on the video.	"Now that you've learned more about giving presentations, can you think of your previous experience and write a comment on how you can improve your oral presentation skills?"	The student tries to reflect on her previous experience or plan for future improvements by considering the main points learned from the video.

3.3.2 Quality Nudges for Example Videos

Since the quality categories for example videos are simple, the nudges only aim at encouraging students to write comments about the oral-presentation components and avoid writing irrelevant comments. Therefore, the first nudge for example videos is the "Elaborate more" nudge which is identical to the one described for tutorial videos. The other nudges for example videos are as follows:

Table 3-10 Definitions of Quality nudges for example videos

Name	Goal	Trigger Conditions	Interaction Template	Outcome
First High-quality Comment Nudge	Encourage the student to reflect on the example video.	The submitted comment is predicted to be in category 2 or 3, and it is the first comment the student has written in category 2 or 3 for the current video.	"Well done, you just made your first reflection on this example video! Keep sharing your thoughts about it."	The student writes more reflections on the example videos.
Frequent High-quality Comment Nudge	Encourage the student to keep writing high-quality comments.	The submitted comment is predicted to be in category 2 or 3, and the student has written more than three comments in category 2 or 3 in the five most recent comments for the current video.	"Great job, you have made a lot of comments criticising this example video. Keep thinking critically about it."	The student develops the habit of writing reflections on the example videos.

3.4 Implementation of Quality Nudges

The existing AVW-Space web application is developed in Python using the Flask[5] web application framework. The data in AVW-Space is stored in an SQLite database using the SQLAlchemy object-relational mapper. The majority of the AVW-Space functionality is accessible via AJAX API requests. There are four components involved in the implementation of nudges: the student model, the nudge engine, the server-side worker, the client-side worker. Figure 3-6 illustrates the interaction of these components. Once a video page is loaded, the client-side (developed in JavaScript) establishes a connection to the server using WebSockets. When the connection is opened, the server creates a worker. The client-side worker pushes state updates to the server worker. Then the server worker updates the student model and notifies the nudge engine on student model updates. The nudge engine evaluates nudges to determine which nudge is appropriate for the current state of the student model. Then, the server-side worker sends the nudge to the client worker. I enhanced the client worker, student model and nudge engine to add the Quality nudges to AVW-Space, on top of existing Reminder nudges.

[5] https://palletsprojects.com/p/flask/

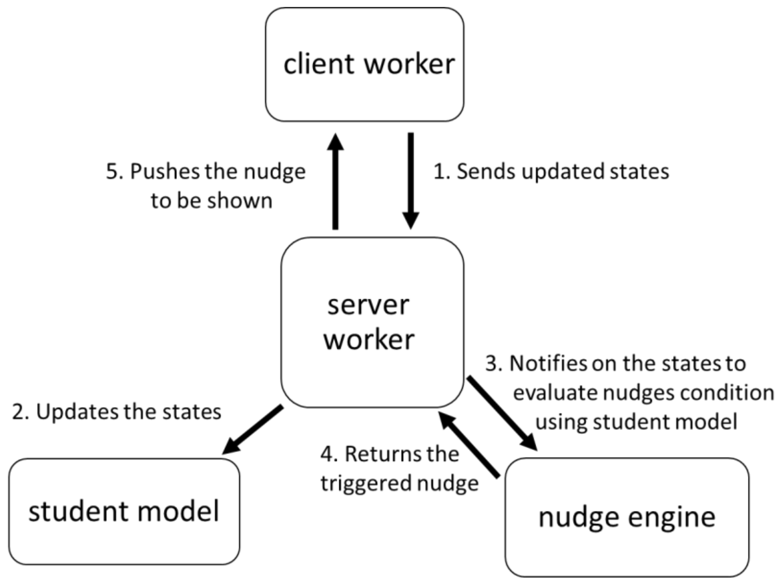

Figure 3-6 Interactions between components in AVW-Space for providing nudges

The student model stores information about the ID and current timestamp of the video that the student is watching, the number of comments the student made on the video and the record of aspects the student used in comments on the video. The student model is updated whenever a video player state changes or a new comment is submitted. The client worker pushes the video player state to the server every two seconds or when the student fast-forwards or rewinds the video. When a student makes a comment, the client worker sends the message and aspect of the comment to the server worker to update the student model. In the enhanced version of AVW-Space, the comment's quality evaluated by the quality assessment models, its domain-specific ratio and LIWC features are also pushed to the server when a comment is submitted. In addition, the comment quality is stored in the comments table in the database.

The nudge engine is implemented as a set of Python classes inherited from a single base class, NudgeBase. Each subclass of NudgeBase can implement the *title()*, *message()*, and *evaluate()* methods. For each Quality nudge, I added a new subclass of NudgeBase and overwrote its methods based on the nudge definition as presented in Table 3-9 and Table 3-10. The *title()* method returns the nudge name; the *message()* method provides the template of the nudge; and the *evaluate()* method evaluates the trigger conditions of the nudge. The *evaluate()* method assesses whether the conditions of the nudge are satisfied by evaluating the student model, the students' MLSQ profiles (if available), the history of nudges that the student received for the video, and the comments' quality made by the student on that video. I added a colour attribute for each nudge to identify the colour of the nudge box displayed to the student.

Previously, the nudges were shown in a grey box, but I defined a colour code for the nudges to improve the user experience and provide an additional visual cue for the students. All positive feedback nudges are shown in a green box. Nudges such as *No Comment*, *Elaborate More* or *Vague Comment* are displayed in a red box since they are more crucial and demand the students' attention. The rest of the nudges are presented in yellow boxes.

Figure 3-7 shows an example of a Quality nudge in AVW-Space. In this example, the student has made several comments which only repeat the content of the video. Therefore, the *Critical Thinking Nudge* is shown to the student.

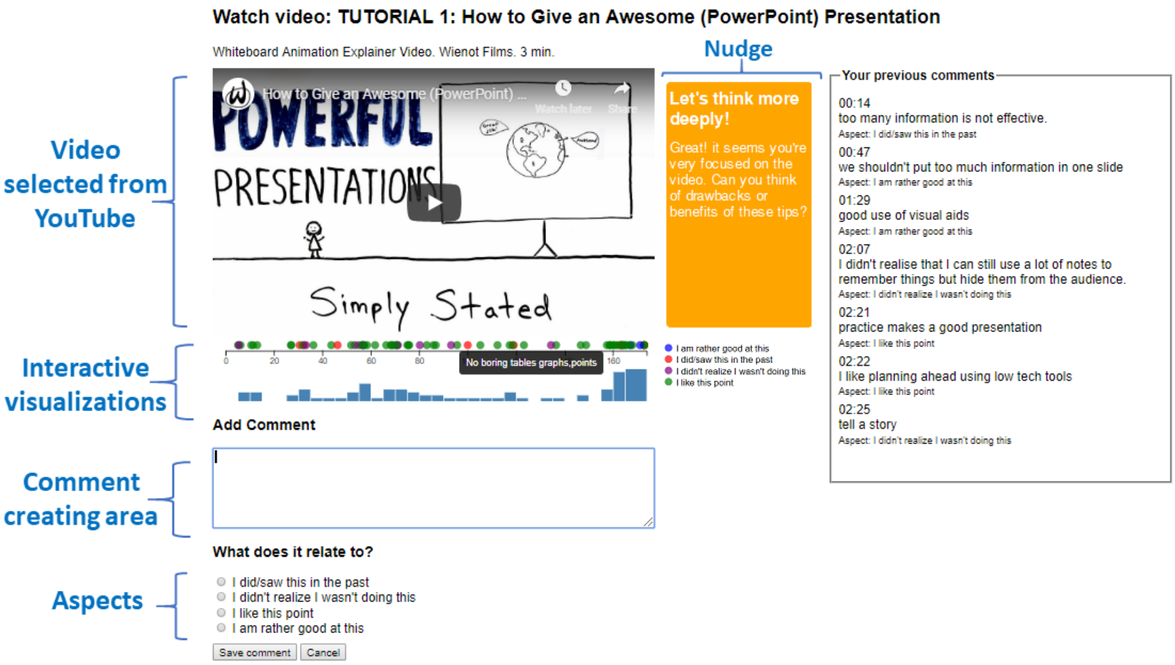

Figure 3-7 An example of a Quality nudge in the enhanced version of AVW-Space

The nudge engine evaluates all nudges in the order of their priority. The priority of the nudges is defined as an attribute of the nudge class. In the enhanced version, nudges such as *No Comment Reminder Nudge* and *No Comment Reference Point Nudge* were prioritised over other nudges since the students must first make comments. Moreover, I defined higher priorities for Quality nudges and lower priorities for *Aspect Under-utilised Nudge* and *Diverse aspects Nudge*. The following is the prioritised list of all AVW-Space nudges:

1. No Comment Reminder Nudge
2. No Comment Reference Point Nudge

3. Final Self-reflection Nudge
4. Elaborate More Nudge
5. Vague Comment Nudge on tutorial videos, First High-quality Comment Nudge and Frequent High-quality Comments Nudge on example videos
6. Repeating Comment Nudge
7. Critical Thinking Nudge
8. First Self-reflection and Frequent Self-reflection Nudge
9. Aspect Under-utilised Nudge
10. Diverse Aspects Nudge

This section addressed RQ3 by explaining the design and implementation of Quality nudges which encourage students to write better quality comments. The effectiveness and usefulness of these nudges were evaluated in the study presented in Chapter 4.

3.5 Conclusions

This chapter presented two quality schemes for assessing comments in AVW-Space and the automation of comments quality assessment by developing cost-sensitive classifiers. The generalisability of these classifiers was assessed on unseen data from two studies with different experimental setups. The performance of the classifiers was slightly lower for the 2017 and PG datasets. Next, the classifiers were used for designing and implementing Quality nudges which encourage students to improve the quality of their comments by triggering critical thinking, self-reflection and self-regulation.

The further improvement in the classifier performance for low-quality comments can be investigated by trying other cost-functions and a more sophisticated feature engineering. Furthermore, since the classifiers were trained and tested only on comments about giving presentations, the generalisability of this approach for other domains should also be investigated in future work. The design of Quality nudges was based on previous studies on learning presentation skills. Therefore, the generalisability of the proposed Quality nudges should be investigated for other soft skills. The existing implementation of the nudge engine allows the addition of new nudges by subclassing NudgeBase. However, the performance of the nudge engine could become unacceptably slow when the number of nudges grows since all nudges are evaluated upon any model update. In addition, choosing the proper priority for each nudge could be a complicated task as the number of nudges grows. Another drawback of the current implementation is that new nudges must be defined in the form of a class. A better solution is creating a database table for nudges to allow teachers to customise the nudges. This table could include the title, message, colour code and priority of the nudge and a set of conditions in a simple formal language to evaluate the nudge.

The Quality nudges require to be tested in studies to evaluate their effectiveness in increasing engagement and learning outcomes and investigate students' attitudes towards these enhancements. The next chapter discusses the study conducted in the same first-year engineering course using the enhanced version of AVW-Space, which includes Reminder and Quality nudges.

4 Evaluating Quality Nudges

To evaluate the effectiveness of *Quality nudges (QN),* I conducted a study with the enhanced version of AVW-Space which included Quality nudges on top of the existing Reminder nudges. I compared the data from this study with previous studies without Quality nudges. In this chapter, I present the effect of QN on learning and engagement, as well as the participants' subjective opinions on QN. Therefore, this chapter addresses the fourth research question of my PhD research:

RQ4. Do Quality nudges increase engagement and learning for students?

In order to investigate this multi-dimensional research question, I defined the following more detailed sub-questions:

RQ4.1. Do Quality nudges increase engagement and foster constructive behaviour? I expected an increase in engagement, as measured by the number of comments and ratings, and the interaction time in the study with QN in comparison to previous studies without QN (Hypothesis H4.1). Since nudges encourage students to write better comments, I expected to see an increase in the quality of comments in the study with QN compared to the studies without QN (Hypothesis H4.2). Another expectation is that there would be more students with constructive behaviour (Hypothesis H4.3), as a result of Quality nudges. I also anticipated a difference in engagement, as measured by the number of videos watched, comments and ratings, interaction time and co-occurrence of the interactions between Constructive students and the other ICAP categories (Hypothesis H4.4).

RQ4.2. What is the effect of Quality nudges on learning? Is there a causal relationship between Quality nudges and students' knowledge? I hypothesised that Quality nudges would increase learning through increasing the number of high-quality comments (Hypothesis H4.5).

RQ4.3. Do students with different engagement levels have different opinions on the usefulness of AVW-Space and cognitive load? Are these opinions different from the ones found in the previous studies? I expected that students with different engagement levels would perceive the effects of nudges differently (Hypothesis H4.6). I also anticipated that there would be a difference in subjective opinions of participants who received Quality nudges in comparison to participants of the previous study who only received Reminder nudges (Hypothesis H4.7).

I also investigated the effectiveness of Quality and Reminder nudges for students who speak English as a Foreign Language (EFL) compared to native English speakers (Native). Video-based learning requires good listening and reading comprehension skills, which could be challenging for EFL students. Moreover, writing high-quality comments in AVW-Space requires good English writing skills. In a study on MOOCs, non-native speakers reported language barriers, such as low reading speed, and cognitive overload and stress related to the visibility of their written responses (Sanchez-Gordon & Luján-Mora, 2014). A follow-up study suggested some approaches for supporting EFL students, such as providing the ability to pause or regulate the video speed, giving access to a downloadable transcript or a domain glossary and using colour and visualisations (Sanchez-Gordon & Luján-Mora, 2015). These approaches are implemented in AVW-Space to some extent (except for the domain glossary). However, adapting learning resources to the student's language competency requires a comprehensive investigation of the characteristics of Native and EFL students. Therefore, after comparing learning strategies of EFL and Native students, I investigated the effectiveness of nudges for EFL and Native students in increasing engagement and learning outcomes to better understand and support EFL students' needs and abilities. Therefore, this chapter also addresses the following research subquestions:

RQ4.4. Do Native and EFL students have different self-reported learning strategies? I first hypothesised that EFL students might have different learning strategies from Native students due to differences in their cultural and educational background (hypothesis H4.8). These differences should be taken into account in the further analysis of their behaviour and learning to design supports for their needs.

RQ4.5. Is there a difference in engagement and learning outcomes for EFL and Native students when receiving nudges? I expected to see a lower engagement in EFL students than in Native students, measured by the number of watched videos, comments (specifically high-quality comments) and ratings made due to language difficulties (hypothesis H4.9). I also anticipated fewer EFL students would show less Constructive behaviour when receiving nudges compared to Native students, since following nudges instructions could be challenging for EFL students (hypothesis H4.10). Finally, I hypothesised that EFL students would have less conceptual knowledge scores due to language difficulties (hypothesis H4.11).

RQ4.6. *What are the differences in the subjective opinions of EFL and Native students on interactions with AVW-Space, when receiving nudges?* I hypothesised that language difficulties might result in EFL students perceiving higher cognitive load in AVW-Space tasks and less usefulness in nudges than Native students (hypothesis H4.12).

In Section 4.1, I describe the experimental design of the study conducted on the Quality nudges. Then, Section 4.2 presents the results from this study. I first address the research questions RQ4.1 to RQ4.3 and investigate the effectiveness of Quality nudges in comparison to the Reminder nudges. Then, I discuss the differences between EFL and Native students and the effectiveness of nudges for these two groups (RQ4.4-RQ4.6). Finally, Section 4.3 summarises the results of this study, discusses the limitations and proposes future work for improving supports for engagement in AVW-Space.

4.1 Experimental Design

I conducted a study on AVW-Space with Quality nudges in the same first-year engineering course (ENGR101) in May 2020. As in previous years, the students were notified about the online training for presentation skills to prepare for the presentation of their final project. The students who watched at least one video on AVW-Space received 1% of the final grade. The ethical approval for the study was obtained from the University of Canterbury in March 2020 (reference ID: HEC 2020/12/LR-PS).

The experimental procedure was identical to the one used in previous studies: participants were invited to complete Survey 1, watch and comment on tutorial videos, and later watch and critique the example videos. During this phase, the participants received personalised nudges (Reminder and Quality nudges). In the second phase, students were instructed to review and rate the comments made by their peers. The comments to be rated were displayed in the order of their predicted quality (from high to low) to facilitate the reviewing task. Finally, Survey 2 was released to students. In Surveys 1 and 2, the participants had one minute to list all concepts they knew about the structure, visual aids, delivery and speech. The students' answers were marked automatically, using the ontology of presentation skills developed in previous work (Dimitrova et al., 2017). The marks for conceptual knowledge questions were used as pre- and post-test scores (CK1 and CK2).

Since this study is a quasi-experiment, it has no control group. Instead, I compared the data from the 2020 study to the data from previous studies conducted in 2018 and 2019, which did not include Quality nudges. As mentioned in Chapter 3, 2018 and 2019 studies had an identical experimental design. In 2018 and 2019 studies, students were randomly assigned to the control or experimental group. Students in the control group did not receive any nudges. In contrast, the experimental group received only the Reminder nudges.

4.2 Results

Out of 947 students enrolled in the course, 490 students interacted with AVW-Space in the 2020 study. 364 students completed Survey 1, and 156 completed both surveys. Table 4-1 presents the demographics of the students who completed Survey 1. There were more male (72.25%) than female participants. The majority of the participants (95.05%) were in the 18-23 age group. Most of the participants (91.01%) were native English speakers. Table 4-1 also presents the mean and standard deviation of the students' responses to the questions about training and experience in giving oral presentations, how often they watch videos on YouTube and how often they use YouTube for learning.

Table 4-1 Demographics of 364 students who completed Survey 1 in the 2020 study

Demographics	Statistical description
Gender	263 male, 99 female, 2 other
Ages 18-23	346
Native English speakers	328
Training	1.73 (.57)
Experience	2.25 (.99)
YouTube	4.35 (.99)
YouTube for learning	3.49 (1.09)

In order to investigate the effectiveness of Quality nudges, I eliminated data for 70 students who did not watch any videos. Then, I compared the remaining data for 294 students with the data from the experimental group of 2018 and 2019 studies, who received only the Reminder nudges. In the experimental group of 2018 and 2019 studies, there were 167 and 171 students respectively, who logged into AVW-Space, completed Survey 1 and watched at least one video. The ANOVA on training, experience and use of YouTube videos in general and for learning showed no significant difference between these three studies. In addition, ANOVA revealed no significant difference in the scores for MSLQ dimensions between the three studies, except in task value ($F = 10.15$, $p < .001$). The 2020 students had significantly higher self-reported scores for task value ($5.63 \pm .80$) than 2018 students ($5.27 \pm .76$). However, the self-reported task value of participants in 2019 ($5.44 \pm .92$) was not significantly different from the other studies.

4.2.1 RQ4.1: Effect of Quality Nudges on Engagement

Table 4-2 presents the statistical description of the activities students performed. ANOVA showed no significant difference in the number of videos students watched between these three studies. However, there was a significant difference in the number of comments made. The post hoc analysis with the Bonferroni correction revealed that the number of comments made in the 2020 study was significantly higher than in the experimental group of 2018 and 2019 studies ($p < .01$). In addition, 2020 participants rated a significantly higher number of comments than the other two studies ($p < .001$). The number of days and sessions in 2020 was significantly greater than in the previous studies. Therefore, Quality nudges increased the duration of the engagement, and hypothesis H4.1 is confirmed. The increase in the number of ratings in 2020 could be attributed to displaying comments for reviewing in the order of high to low quality. Another justification for the increase in the number of ratings could be the improvement in the quality of comments resulting from Quality nudges. This refers to hypothesis H4.2, which is addressed next.

Table 4-2 Summary of activities performed and time spent on AVW-Space (mean and standard deviation)

	2020 (N= 294)	2019 (experimental) (N=171)	2018 (experimental) (N =167)	Significance
Video	6.90 (4.70)	6.49 (3.82)	7.06 (4.21)	$F = .78, p=.45$
Comments	10.50 (14.86)	6.78 (10.43)	6.33 (9.60)	$F = 7.85, p<.001$
Ratings	21.74 (73.39)	3.43 (15.31)	3.86 (20.27)	$F = 9.60, p<.001$
Days	3.08 (1.93)	2.26 (1.81)	2.19 (1.76)	$F = 18.42, p<.001$
Sessions	3.71 (3.17)	2.62 (2.13)	2.53 (2.55)	$F = 13.28, p<.001$

I investigated the performance of the quality assessment models by comparing the comment quality with the human coder labels. The ML classifiers performed well with an F1-score of .83 and .98 for comments on tutorial and example videos, respectively.

For addressing hypothesis H4.2, I compared the distribution of quality categories of comments in different studies (Table 4-3) to investigate the effect of the Quality nudges on the quality of comments. A chi-square test of homogeneity between the studies and quality categories revealed a significant difference (Chi-square = 62.58 and 43.04, respectively, $p < .001$) with effect size (Phi) of .13 ($p < .001$) for comments on tutorial and example videos. I applied a post hoc analysis using the z-test with a Bonferroni correction. The percentage of comments in category 1 on tutorial videos decreased significantly in the 2020 study ($p < .001$). For example videos, there was a significant increase in category 3 comments in the 2020 study ($p < .001$). ANCOVA on the average quality of comments students made on tutorial videos, when controlling for CK1, revealed a statistically significant difference between the 2020 participants (2.04 ± .78) and those of previous years (2018 (experimental): 1.71 ± .10, 2019 (experimental): 1.60 ± .10); ($F = 6.78$, $p < .01$). However, there was no significant difference in the average quality of comments between the experimental group in 2018 and 2019. A similar analysis on the average quality of comments on example videos showed a significant difference in the 2020 study (2.40 ± .02) from the 2018 and 2019 studies (2018 (experimental): 2.27 ± .04, 2019 (experimental): 2.25 ± .04); ($F = 5.50$, $p < .05$). Thus, the quality of comments made by students who received Quality nudges was significantly higher than in previous studies where Quality nudges were not included. Therefore, hypothesis H4.2 is confirmed.

Table 4-3 Distribution of categories of comments in different studies

Video type	Category	2020	2019 (experimental)	2018 (experimental)
Tutorial	Category 1: Affirmative, negative or off-topic	1.3%	6.4%	1.8%
	Category 2: Repeating	47.7%	44.4%	48.9%
	Category 3: Critical and analytical	25.4%	27.7%	23.6%
	Category 4: Self-reflective	23.2%	19.5%	23.1%
	Category 5: Self-regulating	2.5%	2.0%	2.6%
	Total count	1,947	715	653
Example	Category 1: Affirmative, negative or off-topic	1.6%	2.3%	1.5%
	Category 2: Repeating	53.1%	66.2%	66.8%
	Category 3: Critical and analytical	45.4%	31.5%	31.7%
	Total count	1,543	429	394

I categorised the students into three categories using the ICAP framework. Students who watched videos but did not make any comments were classified as Passive. To distinguish Constructive from Active students, I looked at the number of high-quality comments they made on tutorial videos. The median number of high-quality comments on tutorial videos was 2. Therefore, I defined Constructive students as those who wrote three or more high-quality comments, and Active students as those who wrote up to two high-quality comments. Based on this categorisation, there were 75 Passive students, 114 Active students and 105 Constructive students. There was no significant difference between the three categories of students on either demographic or MSLQ dimensions, except for self-reported experience scores ($F = 4.55$, $p < .05$, $\eta^2 = .004$), where Constructive (2.4 ± .71) and Active (2.24 ± .78) students had more experience on giving presentations than Passive students (2.05 ± .78).

I expected to see an increase in the number of Constructive students in the 2020 study, as a consequence of the Quality nudges. Table 4-4 presents the distribution of students over ICAP categories in the three studies. The chi-square test of homogeneity between studies and ICAP categories revealed a significant difference (Chi-square = 26.46, $p < .001$) with effect size (Phi) of .20 ($p < .001$). I applied a post hoc analysis to compare different categories using the z-test with a Bonferroni correction. The percentage of Constructive students increased significantly in the 2020 study ($p < .001$), and the percentage of Passive students decreased significantly in the 2020 study ($p < .005$). Thus, Quality nudges foster constructive behaviour, and hypothesis H4.3 is confirmed.

Table 4-4 Distribution of ICAP categories in different studies

ICAP Categories	2020	2019 (experimental)	2018 (experimental)
Passive	75 (25.5%)	64 (37.4%)	60 (35.9%)
Active	114 (38.8%)	78 (45.6%)	75 (44.9%)
Constructive	105 (35.7%)	29 (17.0%)	32 (19.1%)
Total	294	171	167

I investigated how the different categories of students interacted with AVW-Space in the 2020 study (Table 4-5). The one-way ANOVA tests showed a significant difference for each type of activity between the categories ($p < .001$). Constructive students watched more videos ($F = 33.56$), wrote more comments ($F = 96.09$), received more nudges ($F = 29.06$) and rated more comments ($F = 13.75$) compared to other categories. Passive students made no comments on tutorial videos, as defined earlier. However, two Passive students made comments on example videos which is why the average number of comments in the passive category is not 0. The one-way ANOVA showed that the difference in the percentage of Reminder nudges (RN) and Quality nudges (QN) that each category of students received was significantly different ($F = 36.68$, $p < .001$). Passive and Active students received more Reminder nudges than Quality nudges. However, Constructive students received similar percentages of Reminder nudges and Quality nudges. The number of times Constructive students watched the tutorial videos was significantly higher ($F = 19.65$, $p < .001$) than Active and Passive students, as some Constructive students watched the tutorial videos more than once. Constructive students commented in significantly shorter intervals in comparison to Active students; (F =

20.31, $p < .001$). I also compared the rating options different groups of students used. The most used rating options that Constructive and Active students used were "This is useful for me" (23.93% and 29.73%, respectively) and "I like this point" (54.08% and 47.11%, respectively). However, Passive students used "I hadn't thought of this" (32.50%) and "I like this point" (37.50%) the most but did not use the "I don't agree with this" option at all. This could be because the Passive student did not write any comment in the Personal Space, so they did not develop their own set of opinions to compare with others' opinions.

Table 4-5 Activities performed by different categories of participants

	Passive (75)	Active (114)	Constructive (105)	All (294)
Unique videos	3.8 (2.49)	4.85 (2.79)	6.76 (2.08)	4.68 (3.06)
Times watched tutorial videos	3.42 (2.26)	3.84 (2.08)	5.4 (2.54)	4.29 (2.44)
Comments	.25 (1.05)	6.02 (6.98)	22.68 (17.91)	10.29 (14.78)
Average commenting interval	N/A	252.36 (221.64)	140.81 (96.24)	165.60 (213.14)
Nudges	11.64 (10.50)	17.57 (17.55)	27.96 (13.81)	17.72 (16.33)
RN percentage	.71 (.15)	.60 (.13)	.55 (.07)	.61 (.14)
QN percentage	.28 (.15)	.39 (.13)	.45 (.07)	.38 (.14)
Ratings	5.71 (4.34)	33.63 (81.65)	83.76 (129.02)	19.31 (69.49)

In order to understand how students in different ICAP categories interacted with AVW-Space and used nudges, I applied ENA on the logs of students' interactions in the 2020 study. The logs contain 61,483 entries, which were coded as described in Table 4-6. The video pause and play events do not include auto-pause and auto-play before/after commenting, since I was interested in manual playing/pausing students do intentionally. I defined the codes for comments of specific quality predicted by AVW-Space (CommentQ1/2/3). In addition, I generated specific codes for the Quality nudges, Reminder nudges and nudges presenting positive feedback (*NPositiveFeedback*). Each code is defined for a single action type.

I analysed the actions performed by different ICAP categories on tutorial videos since most students (96.5%) started by watching tutorial videos. As reported in Table 4-7, *VPlay*, *VPause* and *NReminder* are the most frequent events in all three categories. Passive students had the highest frequency of *NReminder* and *VPlay,* as expected, since they only watched the videos. Constructive students received *NPositiveFeedback* more frequently than Active students since Constructive students wrote more high-quality comments (*CommentQ3*). However, the co-occurrence of these actions should be investigated via ENA.

Table 4-6 Description of codes derived from event logs in Personal Space

Code	Description	Example
VPlay	Playing a video	Video state changed: playing
VPause	Pausing a video	Video state changed: paused
VEnd	Reaching the end	Video state changed: page=Watch, state=ended
NReminder	Receiving a Reminder nudge	Nudge: no_comment_reminder_nudge
NQuality	Receiving a Quality nudge	Nudge: final_self-reflection_nudge
NPositiveFeedback	Receiving positive feedback nudge	Nudge: first_self-reflection_nudge
CommentQ1	Making a short/off-topic comment	Comment= “smart“, quality=1
CommentQ2	Making a comment that elaborates or criticises the video	Comment= “Follow a structure and build to conclusion as to make story clear”, quality=2
CommentQ3	Making a reflective or self-regulative comment	Comment= “I try to leave a slide for each key idea”, quality=3

Table 4-7 Frequency of events in tutorial videos for different ICAP categories

Events (14,633)	Constructive (105)		Active (114)		Passive (75)	
	Freq.	Count	Freq.	Count	Freq.	Count
VPlay	32%	2492	36%	1,732	39%	769
VPause	25%	1991	25%	1,218	25%	499
VEnd	4%	278	5%	231	5%	91
NPositiveFeedback	5%	372	2%	101	0	0
NReminder	12%	960	15%	723	21%	410
NQuality	6%	473	8%	380	10%	201
CommentQ1	0	4	0	10	0	0
CommentQ2	11%	884	7%	326	0	0
CommentQ3	5%	407	2%	81	0	0
Total		7,861		4,802		1,970

I generated the epistemic networks using the ENA 1.7.0 Webtool (Marquart et al., 2018), with the units of analysis defined as the ICAP category, subsetted by student_id. I aggregated networks using a binary summation of the co-occurrence of the event codes. The conversation was defined as a single video, subsetted by student_id, and stanza = 3. The reason for this window size is that triggering each nudge involves changes in video status, making a comment and getting the nudge, and I wanted to capture the co-occurrence of these actions. The generated model had co-registration correlations of .97 (Pearson) and .97 (Spearman) for the first dimension of the visualisation space, and co-registration correlations of .97 (Pearson) and .97 (Spearman) for the second. These measures indicate a good fit between the visualisation and the original model.

Figure 4-1 shows the networks generated for each student category. As can be seen, networks generated for Active and Passive students have a strong co-occurrence between *NReminder* and *VPlay* compared to Constructive students, meaning Passive and Active students are more likely to ignore the Reminder nudges and continue watching. In addition, the network generated for Passive students does not have connections to *CommentQ1/2/3* and *NPositiveFeedback* since Passive students did not make any comments.

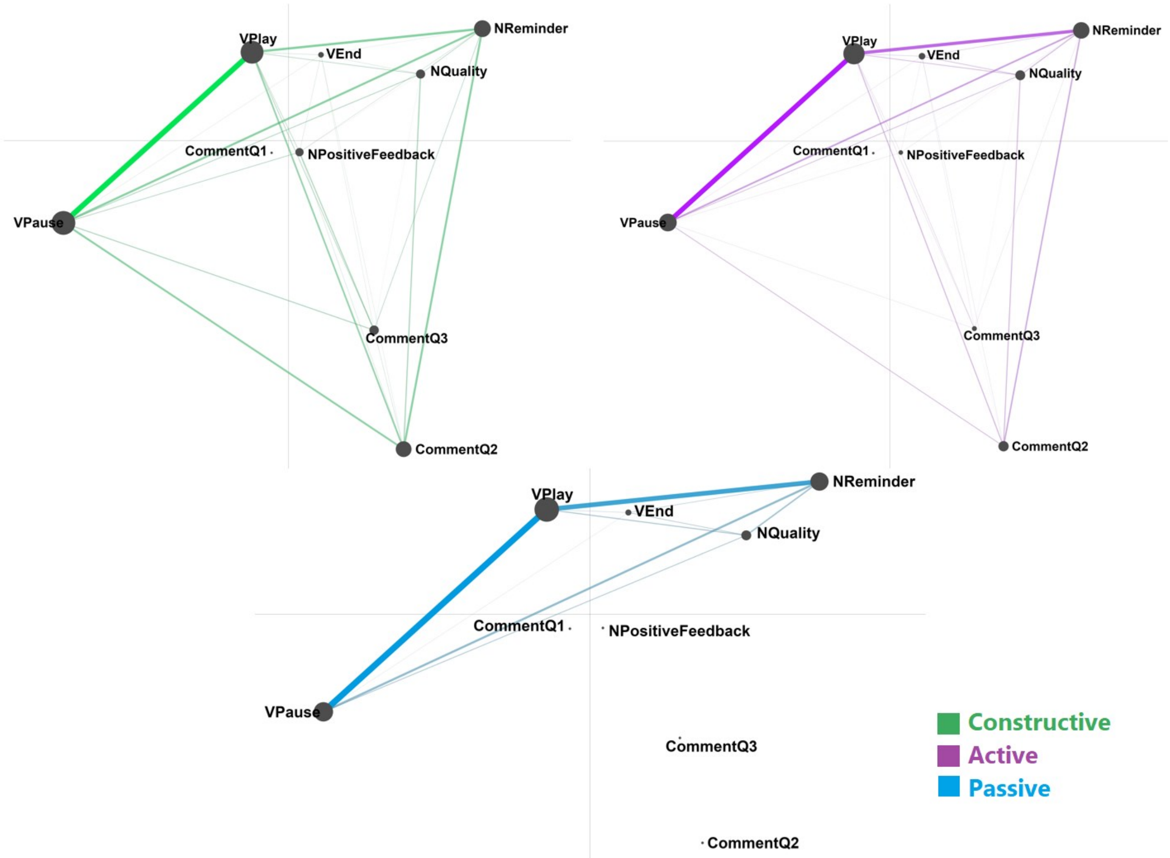

Figure 4-1 Epistemic Networks for responses to nudges for the three student categories

To test the differences between Constructive and Active students (Figure 4-2), a two-sample t-test was applied on the location of points in the projected ENA space. There was no significant difference along the X-axis ($p = .1$) between Active (mean = .13, SD = 1.25) and Constructive students (mean = -.13, SD = 1.02). However, there was a significant difference along the Y-axis ($t = -9.98$, $p < .001$, Cohen's $d=1.36$) between Active (mean = .07, SD = .60) and Constructive students (mean = -.77, SD = .65). The mean for Active students is positioned in the top half of the ENA space, which is more focused on nudges and video watching, while the mean of Constructive students is located in the bottom half, which includes comment making and video pause. The difference network reveals that the co-occurrences of *CommentQ2/3* with *VPause* are stronger for Constructive students than the Active group. Furthermore, the co-occurrences of *VPause* and *CommentsQ2/3* are stronger than nudge-comment co-occurrences for Constructive students. This means Constructive students were more likely to make comments intuitively without being nudged. As expected, co-occurrences of *NPositiveFeedback* and *CommentQ2/3* were more frequent for Constructive students, as they were more likely to continue making good-quality comments. However, the difference network does not have any connection between *NQuality* and *CommentQ3*, meaning there was no difference in this co-occurrence between Constructive and Active students. Thus, Quality nudges were equally helpful for Constructive and Active students to write self-reflective comments. A Quality nudge that students might receive at the end of the video is the *Final Self-reflection Nudge*, which asks the student to make a reflective comment if they have not submitted one yet. The strong connection between *VEnd* and *NQuality* for Active students could indicate that Active students are more likely not to have reflective comments before the end of the video. The stronger connection between *NQuality* and *NReminder* for Active students could mean that Active students were more likely not to satisfy the nudges, so these nudges are repeated more frequently for Active students.

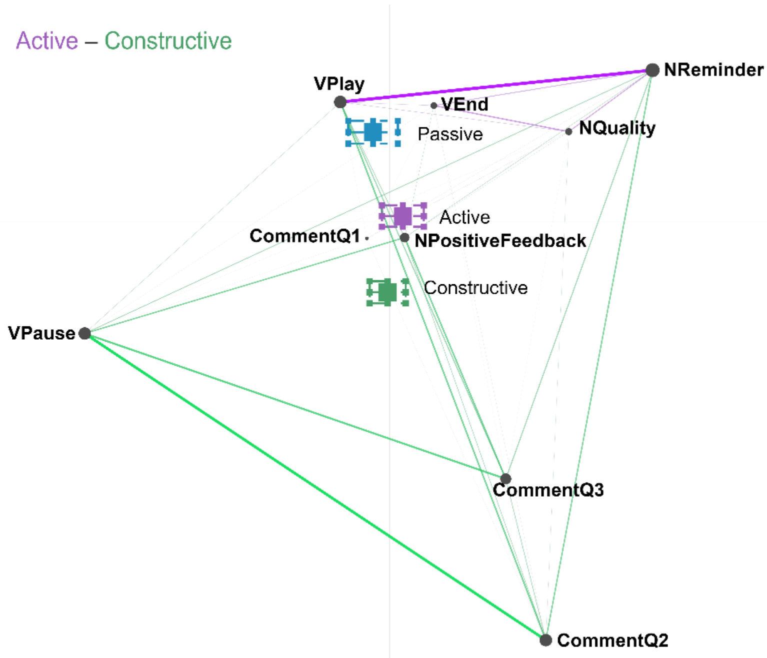

Figure 4-2 Difference network for Constructive and Active students

The two-sample t-test revealed no significant difference between Active (mean = .13, SD = 1.25) and Passive students (mean = -.01, SD = 1.14) on the X-axis (p = .43), but a significant difference along the Y-axis (t = 12.29, p < .001, Cohen's d=1.71) between Active (mean = .07, SD = .60) and Passive students (mean = .98, SD = .42). The difference network for Active and Passive students (Figure 4-3) emphasises the differences in commenting since Passive students did not make any comment. For Passive students, the strong co-occurrence of *NReminder* and *VPlay* is more frequent than the co-occurrence of *NReminder – VPause*, indicating that Passive students were more likely to ignore nudges and continue watching the video rather than pausing the video and reading the nudges and comment. Co-occurrence of *NQuality* and *NReminder* is more frequent for Passive students, indicating Passive students did not satisfy the nudges, and they kept receiving the nudges. The co-occurrences with *VEnd* for Active students indicates that Active students were more likely to watch the video to the end compared to Passive students.

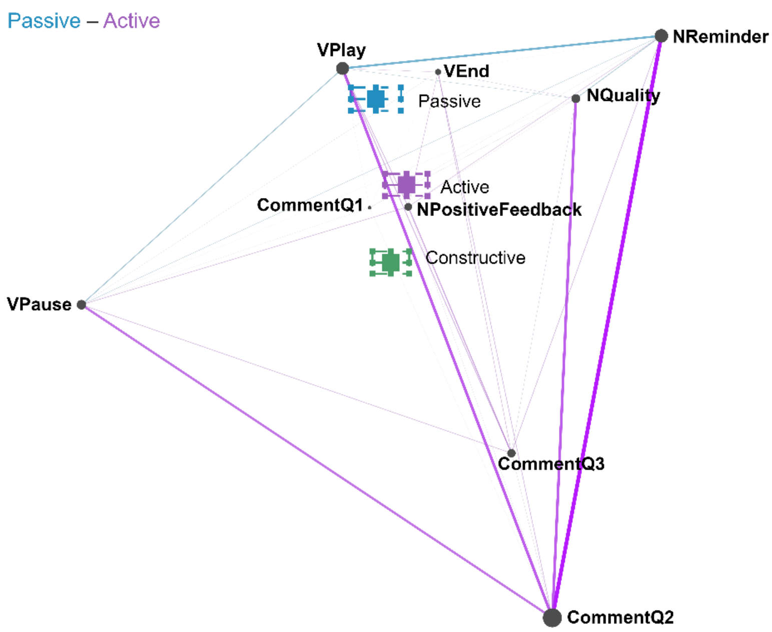

Figure 4-3 Difference network for Active students and Passive students

When comparing Constructive students to Passive students, no significant difference was found along the X-axis (p = .49), but there was a significant difference along the Y-axis (t = 21.95, p < .001, Cohen's d = 3.10) between Constructive students (mean = -.77, SD = .65) and Passive students (mean = .98, SD = .42). The difference network for Constructive and Passive students was similar to the one for Active and Passive students. The ENA networks and results presented in Table 4-5 provide evidence for supporting hypothesis H4.4.

4.2.2 RQ4.2: Effect of Quality Nudges on Learning

I compared the pre-/post-study conceptual knowledge scores of those participants who completed both Surveys 1 and 2 in the three studies to investigate the effects of Quality nudges on learning (Table 4-8). There was no significant difference in CK1 scores from the three studies. I ran an ANCOVA on CK2 scores, with CK1 as a covariate, and found no significant difference between the three studies.

Table 4-8 CK1 and CK2 in different studies

Conceptual knowledge (CK)	2018 (experimental) (N = 102)	2019 (experimental) (N = 131)	2020 (N = 147)
CK1	13.76 (5.51)	13.62 (6.51)	13.14 (5.08)
CK2	14.58 (6.22)	14.88 (6.96)	13.55 (5.71)

Table 4-9 presents the conceptual knowledge scores from Surveys 1 and 2 for different categories of students in the 2020 study. CK1-all is the average CK1 score for all students who completed Survey 1, while CK1 and CK2 are the scores for those participants who completed both surveys. The one-way ANOVA on CK1-all showed a significant difference between the ICAP categories ($F = 8.79$, $p < .001$); Passive-Constructive ($p < .001$) and Active-Constructive ($p = .024$). I ran an ANCOVA to determine the effect of various levels of engagement on CK2 after controlling for CK1, for students in the 2020 study. The assumptions for ANCOVA were met. After applying mean adjustment on CK1, a statistically significant difference was revealed in CK2 for different ICAP categories ($F(3, 144) = 5.40$, $p < .01$, partial $\eta^2 = .07$). The post hoc analysis with the Bonferroni correction showed a significant difference in CK2 for Constructive students compared to Passive students with a mean difference of 3.40 (95% CI, .276 to 6.531) and $p = .02$. There was also a significant difference between Constructive and Active students, with a mean difference of 2.56 (95% CI, .013 to 5.106) and $p < .05$. This result is consistent with previous studies on AVW-Space, where learning outcome was higher for students who wrote more high-quality comments.

Table 4-9 CK1 and CK2 from the 2020 study

Conceptual knowledge (CK)	Passive	Active	Constructive
CK1-all	N = 75 12.38 (5.76)	N = 114 13.49 (5.29)	N = 105 15.64 (5.15)
CK1	N = 30 11.66 (.89)	N = 61 12.48 (5.18)	N= 56 14.66 (4.78)
CK2	N = 30 11.30 (5.34)	N = 61 12.48 (5.03)	N= 56 15.93 (5.85)

I designed a structural equation model to test hypothesis H4.5 and analyse the effect of the number of received Quality nudges on the number of high-quality comments and the conceptual knowledge scores. The model for the 2020 study is illustrated in Figure 4-4. The path diagram consists of rectangles for observed variables, circles for latent variables, curved bidirectional arrows for correlations and straight arrows that link a predicting and a predicted variable. The chi-square test (8.61) for this model (DF = 7, 20 estimated parameters) shows that the model's predictions are not statistically significantly different from the data ($p = .24$). The Comparative Fit Index (CFI) was .989, and the Root Mean Square Error of Approximation (RMSEA) was .039. Therefore, the model is acceptable: CFI is greater than .9, and RMSEA is less than .06 (Hu & Bentler, 1999). The model indicates that the higher CK1 score directly causes a higher CK2 score ($p < .001$). Therefore, the effect of the number of high-quality comments on CK2 is adjusted for and beyond this influence (.51, $p < .001$). All links are significant at $p < .001$ except Quality nudges → High Quality Comments ($p = .053$). Thus, Quality nudges have a positive influence on increasing knowledge. The path diagram and Table 4-9 provide evidence supporting hypothesis H4.5.

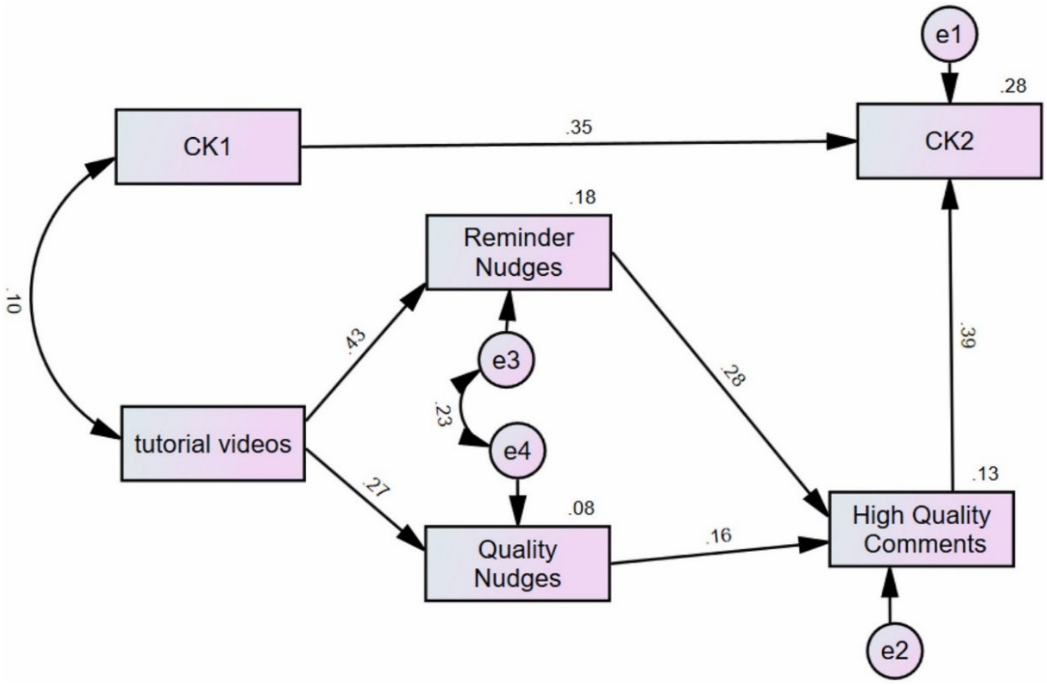

Figure 4-4 Path diagram for testing H4.5: Quality nudges increase the number of high-quality comments and learning

4.2.3 RQ4.3: Subjective Opinions on Quality Nudges

Table 4-10 summarises the responses of different ICAP categories to the TLX-NASA questionnaire in the 2020 study. Constructive students found the commenting task less mentally demanding ($F = 2.93$, $p = .056$) and frustrating ($F = 6.66$, $p < .05$) than other ICAP categories. However, Passive students found commenting more effortful than other students ($F = 6.08$, $p < .01$). Passive students also found comment rating significantly more frustrating ($F = 3.44$, $p < .05$) and more mentally demanding ($F = 5.06$, $p < .05$) than other ICAP categories. This could be because Passive students did not make any comments, and reviewing comments can be difficult for them.

I compared TLX-NASA scores from different studies. The scores for the effort required in commenting (7.71 ± 3.94) were lower in the 2020 study than the experimental group of 2018 and 2019 studies ($F = 2.66$, $p = .07$), which could be a consequence of the Quality nudges in the 2020 study. Moreover, the 2020 participants found the reviewing task more mentally demanding ($F = 4.91$, $p < .01$) than students in previous studies. This could be due to displaying the comments in the order of high-quality to low-quality. However, Table 4-2 showed that the number of ratings in the 2020 study was significantly higher than in previous studies.

Table 4-10 TLX-NASA responses

TLX-NASA	Constructive	Active	Passive
Commenting			
Mental demand	7.89 (3.44)	9.57 (3.79)	9.43 (4.26)
Effort	6.35 (3.22)	8.47 (4.13)	9.06 (4.30)
Frustration	5.37 (4.23)	7.98 (5.45)	9.43 (6.11)
Performance	12.69 (3.13)	11.36 (4.46)	10.36 (4.56)
Comment Rating			
Mental demand	8.10 (3.65)	8.75 (3.79)	10.83 (4.22)
Effort	7.33 (3.55)	8.29 (3.60)	9.00 (4.61)
Frustration	6.69 (5.69)	7.88 (4.90)	9.70 (5.20)
Performance	11.10 (4.36)	10.47 (4.42)	9.43 (4.24)

The scores for the usefulness of AVW-Space in the three studies were similar. However, the 2020 categories differ in their opinions about the usefulness of AVW-Space (Table 4-11). Constructive students are more likely to use AVW-Space frequently ($F = 3.12$, $p < .05$) than Active/Passive students (statement 1 in Table 4-11). Constructive students also found AVW-Space more useful in their studies/jobs than other students ($F = 3.08$, $p < .05$) (statement 6 in Table 4-11). Moreover, Active and Passive students found AVW-Space less easy to do what they wanted, compared to Constructive students ($F = 2.81$, $p < .05$) (statement 7 in Table 4-11). These analyses provide evidence supporting hypothesis H4.6.

Table 4-11 TAM scores in different ICAP groups

TAM statements	Constructive	Active	Passive
1. **I think I would like to use AVW-Space frequently.**	3.66 (1.41)	4.49 (1.66)	3.83 (1.36)
2. I would recommend AVW-Space to my friends.	3.46 (1.41)	4.06 (1.49)	3.8 (1.32)
3. Using AVW-Space would enable me to improve my soft skills quickly.	2.92 (1.42)	3.28 (1.30)	3.36(1.42)
4. Using AVW-Space would improve my performance, considering the development of soft skills.	2.82 (1.44)	3.11 (1.10)	3.1 (1.47)
5. Using AVW-Space would enhance my effectiveness when developing soft skills.	2.80 (1.38)	3.18 (1.04)	3.03 (1.59)
6. **I would find AVW-Space useful in my studies/job.**	2.91 (1.33)	3.58 (1.45)	3.3 (1.53)
7. I would find AVW-Space easy to do what I want it to do.	2.96 (1.40)	3.7 (2.58)	3.33 (1.02)
8. My interaction with AVW-Space would be clear and understandable.	2.87 (1.41)	3.5 (1.37)	3.4 (1.10)
9. I would find AVW-Space easy to use.	2.60 (1.41)	3.2 (1.47)	3.23 (1.16)
10. If I am provided the opportunity, I would continue to use AVW-Space for informal learning.	3.67 (1.66)	4.15 (1.65)	3.9 (1.29)

I collected 185 responses on the usefulness of nudges, strengths and weaknesses of AVW-Space from Survey 2 in the 2020 study. The majority (69.72%) of participants found the nudges helpful for being engaged, writing more comments, directing their thinking and learning while commenting. However, 61.11% / 62.50% of answers to this question found nudges useful in the experimental group of 2018/2019 studies. Two examples of the positive opinions on nudges in the 2020 study were: *"It helped me formulate and structure my response"* and *"They were good as they pointed me in the right direction"*. However, 12.97% of students had negative feedback on nudges in the 2020 study, since they found the nudges distracting or too generic. Two examples of these negative opinions study were: *"Not very useful, I got the idea of the video and they were popuppy and annoying"* and *"The hints had the potential to be helpful, but generally weren't as they were very generic and repetitive"*. The proportion of negative feedback on nudges in the 2020 study was similar to the 2018 study (13.28%), but lower than the 2019 study (20.00%). The rest of the feedback on nudges were neutral. The students' feedback, as well as the comparison of TAM and TLX-NASA scores in the three studies, provide evidence supporting hypothesis H4.7. I also received some suggestions on allowing the students to customise the nudges or improving the user interface of the nudges, for example: *"To adjust the frequency or have the option of turning off reminders"* and *"They were helpful but would be better if you could choose when you saw them. Sometimes they were gone before I understood what they meant."*

Students were also asked to identify the most exciting and disappointing features of AVW-Space in Survey 2. Most students pointed out the videos, rating others' comments and the novelty in the system as the exciting part of AVW-Space in all three studies. Some examples of the responses were: *"Being able to learn about presentations from a range of different perspectives"*, *"It's a very good way to develop your thinking, not just absorb information. It gives critical thinking skills in an effective teaching method"* and *"Learning and discovering what you need to work on"*. However, there were many suggestions for improving the user interface. For instance, some students asked to include a progress bar to visualise their activities: *"It does not monitor your progress. I had no idea how many comments to rate. No good way of recording which videos you've watched or what you've achieved."* These responses reveal several potential improvements to enhance AVW-Space.

The results presented in this section revealed that the Quality nudges effectively enhance constructive behaviour and increase learning outcomes. However, the effectiveness of nudges for students who speak English as a Foreign Language (EFL) could be a concern, since good listening/reading comprehension and writing skills are needed for constructive learning (watching videos, understanding nudges and making high-quality comments). Therefore, further analysis is required for investigating how nudges could benefit EFL students.

4.2.4 RQ4.4: EFL and Native Students' Self-reported Learning Strategies

To compare EFL and Native students, I used the data from studies with AVW-Space conducted in 2018, 2019 and 2020. Due to the small sample size for EFL students in the control and experimental group of 2018 and 2019 studies, I combined the data from groups with similar settings. In other words, I combined the control groups of 2018 and 2019 studies, where no nudge was provided. This section refers to this group as "2018+2019 (control)". I also combined the data from the experimental group of the 2018 and 2019 studies where students received Reminder nudges. I refer to this group as the "2018+2019 (experimental)".

There were 986 students (133 EFL and 853 Native) in all three study groups who completed Survey 1. The first language of most EFL students (44.92%) was Chinese and Vietnamese, while 17.39% spoke European languages (e.g., Dutch and Spanish) and 14.49% Indian languages (e.g., Hindi and Punjabi). Table 4-12 shows the distribution of EFL/Native students in the three study groups and their CK1 scores. There were fewer EFL students in the 2020 study due to the COVID19 travel restrictions. I ran ANOVA on CK1 with study groups and Native/EFL as two factors. The test of between-subject effects showed that the study group had no significant effect on CK1. However, whether students were Native speakers of English or EFL had a significant effect on CK1 ($F = 40.44$, $p < .001$); EFL students had significantly lower CK1 scores compared to Native students. As there are no differences in CK1 scores from the three groups, I combined all Native and EFL students and report analyses done in the following.

Table 4-12 Distribution of EFL/Native students in three study groups and their CK1

Study Groups	#All	#EFL	#Native	EFL CK1	Native CK1
2018+2019 (control)	355	63	292	10.52 (6.48)	12.90 (5.62)
2018+2019 (experiment)	338	42	296	8.19 (5.37)	13.55 (5.86)
2020	294	29	265	11.44 (6.39)	14.25 (5.38)
Total	986	133	853	9.99 (6.25)	13.54 (5.64)

There were no significant differences between Native and EFL students on the self-reported scores for training/experience in giving presentations and using YouTube. Table 4-13 shows the scores on the MSLQ dimensions. Independent t-test showed significant differences in extrinsic goal orientation (t = 4.40, p < .001), rehearsal (t = 4.96, p < .001), self-regulation (t = 2.23, p < .01), organisation (t =2.46, p < .05) and critical thinking (t = 2.99, p < .001). Thus, EFL students reported strong meta-cognitive strategies, but reasons such as grades, rewards, evaluation by others, and competition motivate EFL students more than Native students. Therefore, hypothesis H4.8 is confirmed.

Table 4-13 MSLQ scores for EFL and Native students

MSLQ Score	EFL (N = 133)	Native (N = 986)
Control of learning	5.65 (.85)	5.61 (.77)
Effort regulation	4.73 (1.04)	4.84 (.98)
Elaboration	5.10 (1.06)	5.03 (.89)
Extrinsic goal orientation***	5.80 (.94)	5.38 (1.04)
Intrinsic goal orientation	5.28 (.99)	5.12 (.85)
Self-regulation**	4.68 (.92)	4.42 (.70)
Organisation *	4.89 (1.10)	4.63 (1.12)
Rehearsal***	4.60 (1.16)	4.11 (1.09)
Self-efficacy	5.07 (.96)	5.01 (.93)
Task value	5.43 (.95)	5.49 (.80)
Critical thinking***	4.73 (1.13)	4.35 (1.07)

* p < .05, ** p < .01, *** p < .001

4.2.5 RQ4.5: Engagement and Learning Differences in EFL and Native Students

Table 4-14 compares the number of videos watched, comments and ratings made, and nudges received by EFL and Native students in different studies. I applied ANOVA on these activities, with study group and EFL/Native as two fixed factors. The test of between-subject effects showed that the study group had significant effect on the number of videos (F = 41.93, p < .001), comments (F = 5.49, p < .01), specifically high-quality comments (F = 10.23, p < .001), and ratings (F = 6.48, p < .01), due to the effect of nudges. There was no significant difference in the number of videos watched and ratings made between EFL and Native students. EFL students might find watching English videos less challenging, since AVW-Space allows students to regulate video speed and use YouTube auto-generated subtitles. Besides, reviewing peers' comments can be less challenging than writing comments for EFL students since they become familiar with domain-specific words from watching the videos. However, EFL/Native only had a significant effect on the number of comments (F = 11.17, p < .05). Native students made significantly more comments than EFL students in the 2020 study (F = 3.88, p < .05). In addition, a significant difference was revealed in the number of low-quality ($F=4.53$, $p<.05$) and high-quality comments on tutorial videos ($F=12.63$, $p < .001$) between EFL and Native. Native students made significantly more low-quality and high-quality comments on tutorial videos than EFL students. In the 2020 study, the EFL students received significantly fewer nudges than Native students ($F = 3.95$, $p < .05$). This could be because EFL students made fewer comments, so they received fewer nudges and feedback on their comments. However, there was no significant difference in the number of nudges received by EFL/Native students in the 2018+2019 (experiment) group.

Table 4-14 Statistical descriptions of activities for EFL and Native students

	2018+2019 (control)		2018+2019 (experiment)		2020	
	Native	EFL	Native	EFL	Native	EFL
Videos	6.86 (4.93)	6.77 (4.52)	6.68 (3.96)	7.40(4.38)	7.03 (4.77)	5.72 (3.90)
Nudges	None	None	9.70(10.30)	9.43(5.59)	20.38(16.26)	14.17(12.71)
Comments	4.13 (7.72)	2.90 (6.97)	6.72(10.26)	5.38(8.04)	11.43 (.67)	3.24 (2.02)
Ratings	3.79(18.43)	10.17(46.25)	4.00(19.05)	1.16(4.06)	21.75(72.08)	21.69(85.91)
Low-quality comments	1.80 (3.63)	1.49 (3.03)	2.03 (3.53)	2.02 (3.08)	1.03 (1.88)	2.97 (4.10)
High-quality comments	.80 (1.78)	.30 (.87)	2.06 (3.56)	1.26 (1.98)	3.10 (4.26)	1.13 (2.19)

To better understand the differences between the EFL and Native students' comments, I calculated the average of LIWC features and the domain-specific ratio of comments for each student. There were 273 comments made by 66 EFL students, and 2,318 comments made by 649 Native students on tutorial videos. The independent t-test on the domain-specific ratio showed no significant difference for the comments made by EFL and Native students on tutorial videos. The independent t-test on LIWC features of the tutorial comments showed no significant difference in the comment lengths. However, the t-test showed that the number of words per sentence in comments made by EFL students was significantly lower than for Native students. Table 4-15 shows the mean and standard deviation of LIWC features with significant differences for comments made by EFL and Native students on tutorial videos. EFL students used the first-person singular pronouns ("I", "my", etc.) and auxiliary verbs (such as "will" or "could") significantly less than Native students. Comments showing self-reflection and self-regulation usually contain first-person pronouns (Gašević et al., 2014; Jung & Wise, 2020). Hence, the differences in LIWC scores for EFL and Native students could indicate that the Native students wrote more self-reflective and self-regulating comments. There were no significant differences in cognitive process features such as insight, certainty and differentiation. However, EFL students had a significantly higher score for causation (e.g., "because", "effect", etc.). There were no significant differences in the perceptual process such as seeing, feeling and hearing. However, comments made by EFL students had significantly higher positive emotion scores and significantly lower scores for non-fluent words such as "hm" or "umm".

Table 4-15 Significantly different LIWC features for comments on tutorial videos

LIWC Features	EFL	Native	Significance
First single pronoun	.43 (1.93)	.87 (3.15)	$t = 3.22, p < .01$
Auxiliary verbs	5.33 (8.17)	6.73 (9.14)	$t = 2.64, p < .01$
Causation	4.91 (8.11)	3.70 (6.95)	$t = 2.35, p < .05$
Positive emotions	9.97 (18.68)	7.50 (13.16)	$t = 2.12, p < .05$
Affiliation	1.78 (8.15)	.74 (3.17)	$t = 2.05, p < .05$
Non-fluency	.05 (.55)	.18 (1.75)	$t = 2.61, p < .01$
Word per Sentence	9.61 (7.81)	10.70 (8.40)	$t = 2.06, p < .05$

I also compared the linguistic features of comments made on example videos using the independent t-test (Table 4-16). There were 149 comments made by 51 EFL students and 1,383 comments made by 307 Native students on example videos. Similar to tutorial comments, comments made by EFL students had significantly lower scores for non-fluent words. In addition, comments made by EFL students had lower score in using verbs and adverbs with present focus, but significantly higher domain-specific ratios. This could mean EFL students listed good practices of oral presentation rather than making complete sentences critiquing the presentation in the example video. The results of analysing LIWC features could indicate that EFL students struggle in writing comments showing critical thinking and self-reflection since these types of comments require language proficiency. The results presented in Table 4-14 to Table 4-16 provide evidence for supporting hypothesis H4.9.

Table 4-16 Significantly different LIWC features for comments on example videos

LIWC Features	EFL	Native	Significance
Focus present	8.59 (10.27)	10.99 (11.03)	t = 2.53, p < .05
Non-fluency	.16 (.75)	.87(4.51)	t = 5.18, p < .01
Domain-specific proportion	.34 (.26)	.29 (.24)	t = 2.14, p < .05
Unique Domain-specific proportion	.34 (.26)	.30 (.24)	t = 2.11, p < .05

I investigated the distribution of EFL/Native students in the Passive, Active and Constructive categories (Table 4-17). A chi-square test of homogeneity revealed a significant difference (Chi-square = 16.76, p < .001), with the effect size (Phi) of .13 (p < .001) on all three studies (the *Overall* column in Table 4-17). I applied a post hoc analysis to compare different categories using the z-test with a Bonferroni correction. For EFL students, the proportion of the Constructive category was significantly lower (p < .05) than other categories, while for Native students, the proportions of the different categories were similar. It can be seen that the majority of EFL students were Passive, which indicates the need to provide more focused support for them.

Table 4-17 Distribution of EFL/Native students in ICAP Categories for different studies

		2018+2019 (control)	2018+2019 (experiment)	2020	Overall
EFL	Passive	34 (54.0%)	19 (45.2%)	15 (51.7%)	68 (50.7%)
	Active	26 (41.3%)	15 (35.7%)	9 (31.0%)	50 (37.3%)
	Constructive	3 (4.8%)	8 (19.0%)	5 (17.2%)	16 (11.9%)
Native	Passive	161 (55.1%)	108 (36.5%)	60 (22.6%)	329 (38.6%)
	Active	95 (32.5%)	107 (36.1%)	105 (39.6%)	307 (36.0%)
	Constructive	36 (12.3%)	81 (27.4%)	100 (37.7%)	217 (25.4%)

I also investigated the effect of nudges on EFL/Native students' engagement. A chi-square test of homogeneity between the study groups and the ICAP categories of EFL students revealed a significant difference (Chi-square = 6.16, p < .05) with the effect size (Phi) = .21 (p < .05). Adding Reminder nudges in the 2018+2019 (experiment) group raised the percentage of Constructive EFL students significantly compared to the 2018+2019 (control) group (p < .05). However, the percentage of Constructive EFL students was not significantly different between the 2018+2019 (experiment) group and the 2020 study. Moreover, there was no significant difference in the proportion of Passive EFL students between the study groups.

A chi-square test of homogeneity between the study groups and ICAP categories for Native students showed a significant difference (Chi-square = 76.39, $p < .001$) with the effect size (Phi) of .29 ($p < .001$). The percentage of Constructive Native students increased significantly by including the Quality nudges in the 2020 study, compared to the 2018+2019 (control) group who received no nudge, and the 2018+2019 (experiment) group who received only Reminder nudges. Unlike EFL students, the percentage of Passive students decreased significantly by providing the Reminder nudges in the 2018+2019 (experiment) group and adding Quality nudges in the 2020 study. Thus, the nudges were more effective for Native students than EFL students, and hypothesis H4.10 is confirmed. This could also indicate that the model which triggers Quality nudges is tailored to the behaviour of Native students more than EFL students.

I compared the CK2 of EFL and Native students to determine whether there was a difference in learning. Since only 622 students completed Survey 2, CK1 and CK2 scores were available only for 80 EFL students and 542 Native students. I ran an ANCOVA on CK2 scores, with CK1 as a covariate, and group study and being EFL/Native as two fixed factors. No significant difference was revealed in learning between different study groups. Applying the mean adjustment on CK2 using Bonferroni correction showed that EFL students learned significantly less ($12.17 \pm .64$) than Native students ($14.37 \pm .23$); ($F = 10.37$, $p < .001$). However, lower CK1 and CK2 scores in EFL students could be due to language barriers that EFL students might struggle with in answering the conceptual knowledge questions in Surveys 1 and 2, while they might have learned the skill.

To find the factors influencing learning for EFL/Native students, I ran a generalised linear regression (GLM). For this model, I used CK1 and the number of comments made, with being EFL/Native as the fixed factor to predict CK2. This model was applied only to the 2020 study. The model fitted with Akaike's Information Criterion (AIC) = 3,793.12 (Table 4-18). CK1 and the number of comments were both significant predictors for EFL and Native students. However, each additional point on CK1 has a .15 extra effect on CK2 for Native students (the interaction effect of CK1*Native is .15). The negative coefficient for CK1*Comment shows that the effect of CK1 on CK2 decreases with the rise in the number of comments. Hence, hypothesis H4.11 is confirmed.

Table 4-18 Significant predictors of CK2 for EFL/Native students

Variables	Coefficient	Significance
Intercept	4.96	p < .001
CK1	.50	p < .001
CK1*Native	.15	p < .005
Comment	.21	p < .001
CK1*Comment	-.007	p < .05

4.2.6 RQ4.6: Subjective Opinions of EFL and Native Students

I investigated the responses of EFL/Native students to the NASA-TLX and TAM questionnaires. There was no significant difference in the perceived mental demand, required effort and confidence in performance for the two tasks between Native/EFL students. However, EFL students found the rating task significantly more frustrating (8.81 ± .65) than Native students (7.30 ± .24); (F = 4.70, p < .05). EFL students also perceived frustration during commenting (8.85 ± .66) significantly more than Native students (7.43 ± .24); (F = 4.08, p < .05). Students had no significantly different opinions on the usefulness of AVW-Space, except that the EFL students had lower scores (3.55± .19) for "*I think I would like to use AVW-Space frequently*" (F = 9.65, p < .01) and "*If I am provided the opportunity, I would continue to use AVW-Space for informal learning*" (3.50 ± .19, F = 5.60, p < .05) compared to Native students (4.17 ± .07 and 4.00 ± .07, respectively).

I looked at EFL students' feedback on nudges in the 2020 study. Some EFL students reported that the nudges distracted them from videos (e.g., *"not very useful, took away from the video"*) or they were not confident in writing comments (e.g., *"I am not confident"*, "*They were not useful since I did not know what to do to start with"*). In addition, some responses from passive EFL students showed that they misunderstood the purpose of nudges, such as: *"[nudges helped me] to understand some features I did not know*". There was also some positive feedback from passive EFL students, reporting nudges were useful (e.g., "*Give me directions*", "*Somewhat helpful to remind the user to write a comment*"). However, given that these students were in the Passive category, the nudges were not effective enough for encouraging them to make comments. The EFL students' feedback on nudges and comparison of TAM and TLX-NASA scores in the three studies provide evidence supporting hypothesis H4.12.

4.3 Conclusions

In this Chapter, I investigated the effectiveness of Quality nudges by presenting a study with the enhanced version of AVW-Space, which included Quality nudges. The study revealed that Quality nudges effectively enhanced engagement; participants in this study wrote more comments and of better quality. The Quality nudges caused students to spend more time with AVW-Space. This study showed that the students who received Quality nudges also rated more comments written by their peers. ENA on various ICAP categories' interactions revealed significant differences in the use of nudges. Passive students ignored nudges and continued watching videos rather than commenting. On the other hand, Active students were likely to make no reflective comments until receiving the final self-reflection nudge at the end of the video. Although Constructive students received more nudges, they tended to be less reliant on the nudges to make reflective comments. This study also showed an increase in the conceptual knowledge of participants who made more high-quality comments and received more Quality nudges significantly. This result confirms the findings of the literature on high learning outcomes of constructive students who show more self-reflective and self-regulating behaviours. This group of students also perceived less effort required in commenting and found rating tasks less frustrating and mentally demanding.

I compared the effectiveness of nudges for EFL students with Native students since language difficulties can hinder the effectiveness of commenting and receiving nudges in video-based learning. I investigated the differences between EFL and Native students in their learning strategies, engagement and learning outcomes in AVW-Space. I found that most EFL students watched videos without writing comments, even after receiving Reminder and Quality nudges. Furthermore, EFL students had lower conceptual knowledge scores before and after the study compared to Native students. Although adding Reminder nudges increased constructive behaviour in EFL students, including Quality nudges was not as effective for EFL students as for Native students. The analysis of comments showed significantly fewer indicators of self-reflection in comments made by EFL students than Native students. The generalised linear model also revealed the importance of commenting for EFL/Native students. The comparison of subjective opinions of the EFL student showed confusion about nudges, lack of confidence in making comments and frustration in commenting and reviewing tasks. These insights encourage us to provide more support to help EFL students benefit from video-based learning as much as Native students.

One of the limitations of this study is that the population were only from an engineering background. The enhanced version of AVW-Space with Reminder and Quality nudges needs to be studied on students from a non-engineering background. Another limitation of this research is the low percentage of EFL students in the study population. The main challenge of this research is quantifying learning of a transferable skill like oral presentation. I measured learning by the number of domain-specific phrases that students listed in the pre/post-study surveys. This type of assessment is memory-based and does not fully represent the learner's competency in transferable skills. On the other hand, this approach requires English competency, which could be challenging for EFL students. One approach is observing students giving presentations. However, this approach was not practical in our study since the course population was large, and students were presenting in teams. Hence, further research is required to choose a more comprehensive method for assessing learning outcomes.

This study focused on the effectiveness of Quality nudges in the context of learning oral presentation skills. Therefore, the generalisability of the Quality nudges for other domains should be investigated in future work. The Quality nudges could be improved further by considering those students who did not respond to nudges. For example, when a student consistently makes comments that repeat the video content, giving a general nudge for critical thinking would not be helpful. Instead, the student might require more specific instruction about thinking critically. In addition, the identified significant differences between EFL and Native students could lead to specifying tailored nudges for EFL students. Since nudges are visible for a short period (20 s), it could be useful to visualise the previously received nudges to help students review and reflect on their performance. This visualisation could also be helpful for EFL students to have enough time for comprehending the received nudges. Moreover, the literature suggested that providing information and feedback using colours and signs could reduce the cognitive and information load for EFL students. Since EFL students reported higher motivation towards extrinsic goals, they could also benefit from a dashboard that visualises their progress or allows them to compare themselves with the class. Providing a progress report could also remind all students of the importance of each activity in AVW-Space, such as commenting or fully watching videos. Adding a glossary of main concepts could also help EFL students understand videos and improve their vocabulary. The next chapter presents the implementation of the suggested visualisations and investigates their effectiveness in the engagement and learning of students.

5 Enhancing Visualisations and Evaluating their Effectiveness

Providing visualisation of students' performance could increase engagement in VBL. Previous chapters emphasised the limited research on visual learning analytics (VLA) in VBL and discussed the potential visualisations that could support engagement in AVW-Space. This chapter first presents the integration of various visual learning analytics into AVW-Space, on top of the Quality and Reminder nudges. Then, I discuss the effectiveness of these visualisations in a study conducted on the enhanced version of AVW-Space to address my fifth PhD research question:

RQ5. Does the visualisation of the student model increase engagement and learning? Similar to the previous chapter, I investigated this question by addressing the following subquestions:

RQ5.1. Do the visualisations increase engagement and foster constructive behaviour? Since visualisations show students' progress in learning activities, I expected to see an increase in the number of watched videos, comments, rated videos, diversity of used rating options, and interaction duration in the study with the new visualisations compared to the previous study (2020) which did not offer these visualisations (Hypothesis H5.1). I also anticipated more students with constructive behaviour resulting from the new visualisations (Hypothesis H5.2). In addition, I hypothesised that there would be a difference in engagement, as measured by the number of videos watched, comments, ratings and interaction with visualisations between constructive students and the other categories (Hypothesis H5.3).

RQ5.2. What is the effect of the visualisations on learning? I expected a causal relationship between interactions with visualisations and students' knowledge. As was the case in previous studies, I hypothesised that visualisations would increase learning by increasing the number of high-quality comments (Hypothesis H5.4).

RQ5.3. Do students who were provided with the visualisations have different opinions on AVW-Space's usefulness and cognitive load from participants in the previous study without the visualisations? I expected that participants provided with the visualisations would find visualisations useful and perceive less cognitive demand in interacting with AVW-Space than participants of the previous study who did not receive the visualisations (Hypothesis H5.5).

In the previous chapter, the comparison of EFL and Native students' learning strategies, engagement and learning outcomes suggested providing visualisations to reduce cognitive load for EFL students. Therefore, I briefly investigated the engagement and learning outcome of EFL and Native students in the study with the visualisations to address the following question:

RQ5.4. Does the engagement and learning of EFL students increase when the visualisations are provided? Previous studies on nudges showed that most EFL students were in the Passive category, due to language difficulties, even after receiving nudges. Therefore, I expected to see an increase in the number of Constructive EFL students, as well as the increase in learning after receiving visualisations compared to the previous study without the new visualisations (hypothesis H5.6).

This chapter first presents the design and integration of the new visualisations in Section 5.1. Then, Section 5.2 describes the evaluation study conducted on AVW-Space with the visualisations. Next, Section 5.3 investigates the effectiveness of visualisations and addresses RQ5.1-RQ5.4 by comparing this study with the previous study without the visualisations. Finally, Section 5.4 discusses the results of this study and its challenges, followed by suggestions for future improvement of visualisations.

5.1 Enhancing Visualisations in AVW-Space

In order to decide what visualisations to integrate into AVW-Space, I investigated the students' feedback from previous studies. The majority of students asked for a progress visualisation to monitor what videos they had watched and reviewed. Some students also complained that the nudges disappear before they read them thoroughly, so they would like to revisit the nudges they received previously. In addition, students wanted to see what ratings they had received from their peers in the Social Space. In addition, I noticed that the comment histogram visualisation does not convey additional information to the student on top of the comments timeline. Therefore, I decided to provide a more personalised visualisation for students to compare their comment timeline to the others' comment timeline. I conducted rapid prototyping and evaluated each iteration of prototypes by brainstorming and interviewing with five domain experts to receive feedback. The visualisations went through three iterations: one paper-based mock-up, one digital mock-up and functional visualisation developed using D3.js[6] and JavaScript.

Figure 5-1 shows a screenshot of the new version of AVW-Space, which includes the progress visualisation at the top of the page, the list of videos (not changed from the previous version) and the new green tick icons indicating which videos were visited in the Personal and Social Space. Each student can only see their own progress report. The progress report presents the number of watched videos, commented videos, and videos on which comments are reviewed and rated. The progress report also includes two circles indicating whether Surveys 1 and 2 are completed. The tasks in the progress report are presented in the preferred order: watch a video, comment, and rate peers' comments.

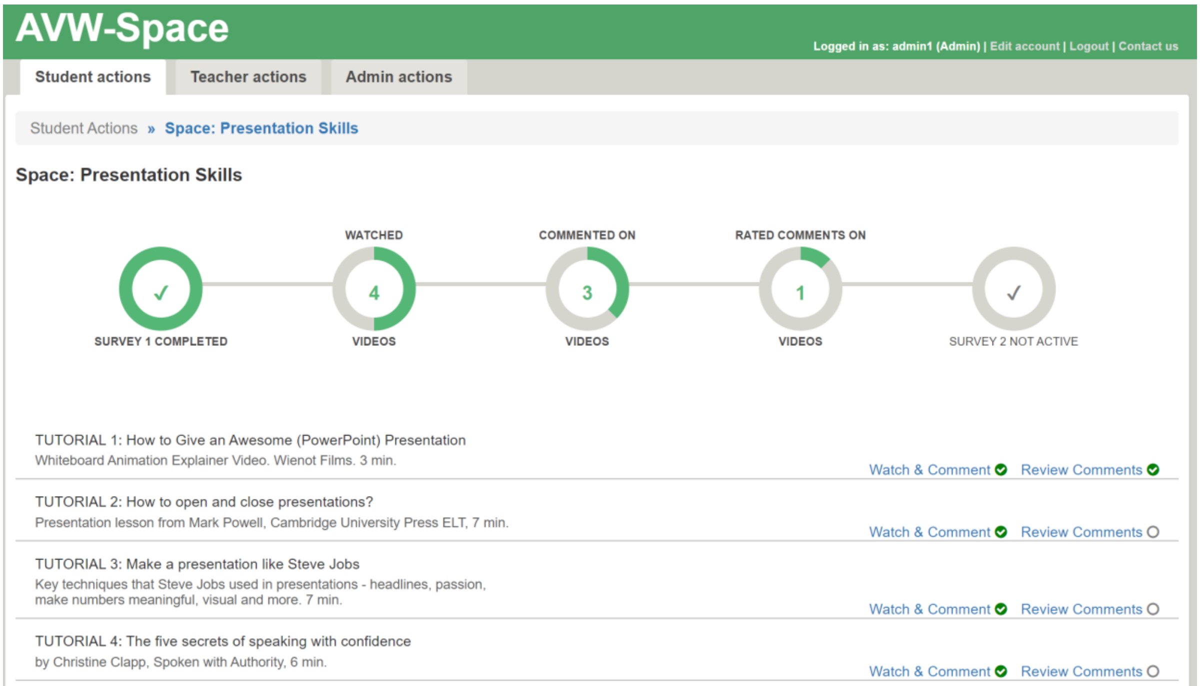

Figure 5-1 Enhanced videos page with progress visualisation

I enhanced the Personal Space, as shown in Figure 5-2. I replaced the histogram visualisation with a visualisation showing the student's own comments to allow comparing their commenting behaviour with the previous class. In addition, the student's comments list on the right side presents the nudges the student received, and a quality indicator for each comment the student wrote. Students can click on the dialogue icon to see what nudge they received at a particular time in the video. The students can hover over the quality indicator to

[6] https://d3js.org/

read the scores for their comments. The quality indicators are in three colours: red (predicted as category 1: off-topic), yellow (predicted as category 2+3: reflecting on the video content) and green (predicted as categories 4+5: self-reflective or self-regulating).

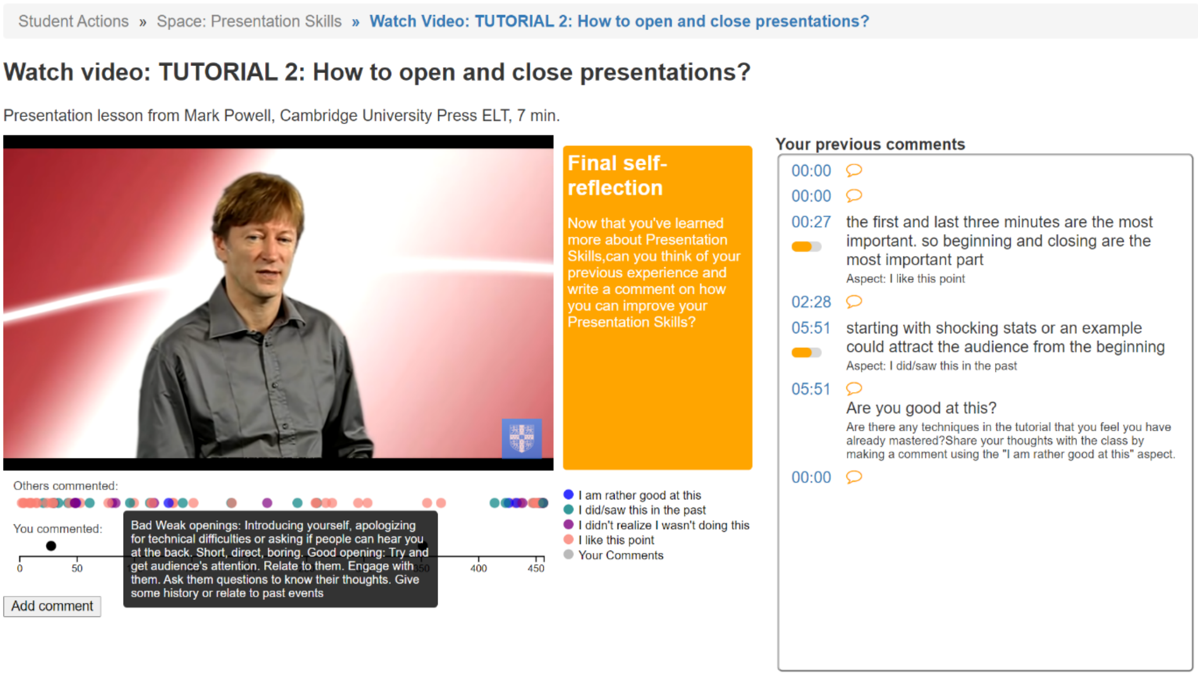

Figure 5-2 Enhanced Personal Space interface with the new visualisations

I also modified the interface for rating comments in Social Space (Figure 5-3). The student can now see a pie chart for their own comments. The pie chart shows ratings the student received from others (anonymised) for each comment. As shown in Figure 5-3, when the student hovers over a rating option on the pie chart, the number of ratings received for that particular rating option will be shown to the learner. However, each student can see only ratings of their own comments. Since the list of comments is sorted by the video timestamp and their quality, the students might find it hard to look for their own comments and the visualisation of their received ratings. Thus, I provided a toggle switch that allows students to see their comments first and then their peers' comments.

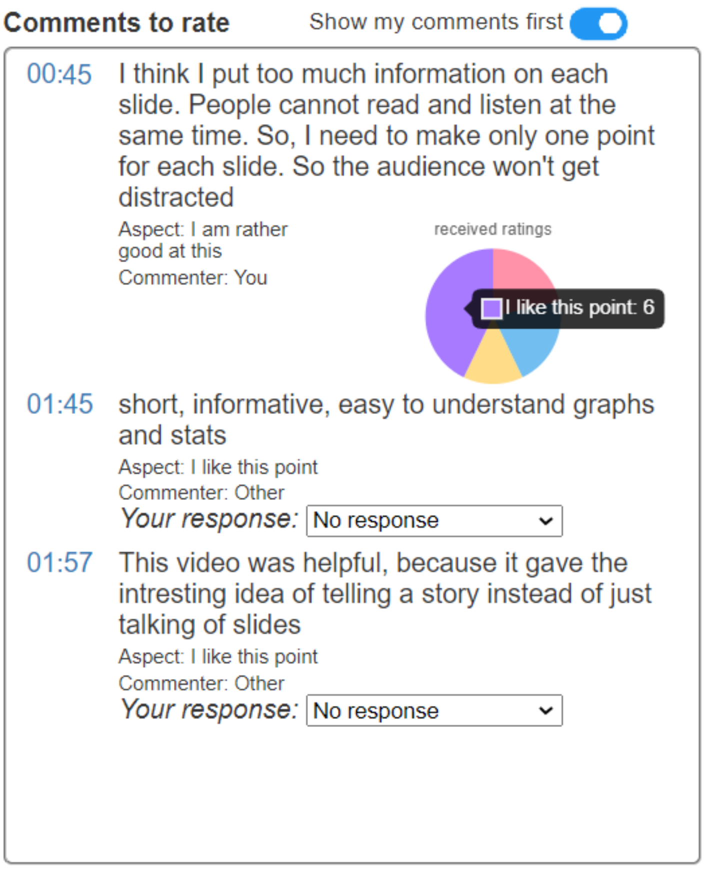

Figure 5-3 Enhanced interface for rating comments in Social Space

5.2 Experimental Design

I conducted a study in the same first-year engineering course (ENGR101) in May 2021. As in previous studies, the students were notified about the online training for presentation skills to prepare for the presentation of their final project. The students who watched at least one video on AVW-Space received 1% of the final course grade. The ethical approval was obtained from the University of Canterbury in April 2021 (reference ID: HEC 2020/12/LR-PS Amendment 1).

The experimental procedure was identical to previous studies: participants were invited to complete Survey 1. Then, participants were instructed to watch and comment on the tutorial videos first, critique the example videos, and finally rate others' comments. Participants in the 2021 study were able to see the new visualisations, in addition to the Reminder and Quality nudges. At the end of the study, participants completed Survey 2, which contained the same questions as in previous studies, along with additional questions on the students' perceived usefulness of each visualisation. The pre- and post-study conceptual knowledge scores (CK1 and CK2) were calculated similarly to previous studies.

Since this study has a quasi-experimental design, I evaluated the effects of the new visualisations by comparing the data from the 2021 study to the data from the 2020 study, in which AVW-Space included only the Reminder and Quality nudges and none of the new visualisations.

5.3 Results

Table 5-1 shows the participation of students in the 2020 and 2021 studies. Similar percentages of students completed Survey 1 and interacted with AVW-Space in both studies. However, the percentage of students who responded to both surveys and interacted with the system (i.e., watched at least one video) increased by 11% in 2021. This could be due to the inclusion of survey completion status in the progress visualisation.

Table 5-1 Participation of students in 2020 and 2021 studies

	2021	2020
Logged into AVW-Space	604 (57.96 %)	490 (52.01 %)
Completed Survey 1	412 (39.53 %)	364 (38.64 %)
Completed Survey 1 and watched videos	351 (33.68 %)	294 (31.12%)
Completed the study	277 (26.58 %)	147 (15.60 %)

The analyses provided in this chapter focus on the students who completed Survey 1 and interacted with the system. Table 5-2 presents the demographics of the students in the 2021 study. Like the 2020 study, there were more male than female participants in the 2021 study. The majority of the participants (94%) were in the 18-23 age group, and most participants were native English speakers. An independent t-test between the 2020 and 2021 studies showed no significant difference in the students' self-reported scores for training and experience in giving oral presentations, how often they watch videos on YouTube and how often they use YouTube for learning. Moreover, there was no significant difference in CK1 ($t = .90$, $p = .36$).

Table 5-2 Demographics of students in the 2021 study

Demographics	2021 study (N=351)
Gender	244 male, 106 female, 1 other
Ages 18-23	329
Native English speakers	308
Training	1.64 (.60)
Experience	2.18 (.77)
YouTube	4.17 (1.07)
YouTube for learning	3.35 (1.06)
CK1	14.18 (.32)

5.3.1 RQ5.1: Effects of Visualisations on Engagement

To address H5.1, I investigated how the students interacted with AVW-Space in the 2020 and 2021 studies (Table 5-3). The 2021 students watched more videos and wrote more comments than the 2020 students. However, there was no significant difference in the number of ratings made. Although there was no significant difference in the number of Reminder nudges, the 2021 students made significantly more comments than students in the 2020 study. However, the 2021 students received significantly more Quality nudges on their comments compared to the 2020 students. The increase in the number of comments could result from the visualisations provided in the Personal Space. Moreover, the 2021 participants commented on significantly a greater number of videos in comparison to 2020. The number of videos in which their comments were rated in the 2021 study was six times more than the 2020 study. Therefore, the progress visualisation helped students complete commenting and rating tasks for more videos. The t-test also showed that the number of days spent on AVW-Space in 2021 was significantly greater than the previous year. Therefore, the students who received the visualisations had a longer engagement duration.

Table 5-3 Activities (mean, standard deviation) in the 2020/2021 studies

	2020 (N = 294)	2021 (N = 351)	t-test
Unique videos	5.26 (2.74)	6.98 (2.24)	$t = 8.54, p < .001$
Comments	10.29 (14.78)	14.04 (11.43)	$t = 3.34, p < .001$
Reminder nudges	11.68 (9.43)	11.90 (6.08)	$t = 344, p = .73$
Quality nudges	8.13 (8.91)	11.36 (8.50)	$t = 4.68, p < .00$
Ratings	21.74 (73.39)	23.55 (52.26)	$t = .36, p = .71$
Videos commented	3.78 (3.29)	6.60 (2.74)	$t= 11.67, p < .001$
Videos rated	1.16 (2.31)	6.44 (3.06)	$t = 24.91, p < .001$
Days spent on AVW-Space	3.08 (1.93)	4.29 (2.93)	$t = 5.73, p < .001$

I also investigated the distribution of rating options used in both studies (Table 5-4). A chi-square test of homogeneity between studies and rating options revealed significant differences (Chi-square = 467.21, $p < .001$) with a significant effect size of Phi = .164, $p < .001$. I applied a post hoc analysis to compare the proportion of rating options using the z-test with a Bonferroni correction. As shown in Table 5-4, the students who were provided with visualisations in the 2021 study used the "I like this point" option less often than the 2020 students. However, the 2021 participants used more reflective options such as "This is useful for me", "I didn't notice this" and "I hadn't thought of this" than the 2020 participants. Therefore, students who received rating visualisations used more diverse rating options. The analyses provided in Table 5-3 and Table 5-4 confirm hypothesis H5.1.

Table 5-4 Distribution of rating options

Ratings	2021 (10,289)	2020 (7,040)	Significance
This is useful for me	3,832 (37.2%)	1,794 (25.5%)	$p < .001$
I hadn't thought of this	1,330 (12.9%)	735 (10.4%)	$p < .001$
I didn't notice this	821 (8.0%)	458 (6.5%)	$p < .001$
I don't agree with this	604 (5.9%)	388 (5.5%)	$p = .31$
I like this point	3,702 (36%)	3,665 (52.1%)	$p < .001$

I categorised the students post hoc into three categories, similar to the previous study, using the ICAP framework. Students who watched videos but did not comment on tutorial videos were classified as Passive. To distinguish Constructive from Active students, I looked at the number of high-quality comments they made on tutorial videos. The median number of high-quality comments on tutorial videos was 2 in both 2020 and 2021 studies. Therefore, I defined Constructive students as those who wrote three or more high-quality comments and Active students as those who wrote up to two high-quality comments. Table 5-5 presents the distribution of students over ICAP categories in the 2020 and 2021 studies. The chi-square test of homogeneity between studies and ICAP categories revealed a significant difference (Chi-square = 45.24, $p < .001$) with effect size (Phi) of .26 ($p < .001$). I applied a post hoc analysis to compare the proportion of different categories using the z-test with a Bonferroni correction. The percentage of Constructive students increased significantly in the 2021 study ($p < .001$), while the percentage of Passive students decreased significantly ($p < .001$). Thus, the new visualisations foster constructive behaviour, and hypothesis H5.2 is confirmed.

Table 5-5 Distribution of ICAP categories in the 2020 and 2021 studies

ICAP Categories	2021 (N = 351)	2020 (N = 294)	Significance
Passive	25 (7.1%)	75 (25.5%)	p < .001
Active	141 (40.2%)	114 (38.8%)	p = .68
Constructive	185 (52.7%)	105 (35.7%)	p < .001

To address hypothesis H5.3 (i.e., there would be a significant difference in engagement), I investigated the interactions of three different categories of students in the 2021 study. Table 5-6 presents the number of videos watched, received nudges, comments, ratings, interactions (hovering for longer than 5 seconds or clicking) with various visualisations, and the number of days and sessions spent on AVW-Space for each category of students. The ANOVA on the activities with student category as the fixed factor showed a significant difference for each type of activity between the categories, except interaction with nudge visualisations. The post hoc analysis with the Bonferroni correction showed that Constructive students watched significantly more videos ($p < .01$), received more Reminder nudges and Quality nudges ($p < .01$), made more comments ($p < .001$) and ratings ($p < .05$) than Active students. In addition, the Active group had significantly more of these interactions than the Passive group ($p < .01$) except in ratings. There was no significant difference in the number of times Active and Constructive students watched tutorial and example videos. However, the number of times Passive students watched the tutorial and example videos were significantly less than Active and Constructive students ($p < .001$).

Constructive students interacted with all visualisations significantly more than Active students, except nudge visualisations. Constructive students interacted with the progress visualisations significantly more than Active students ($p < .05$), but there was no significant difference in interaction with progress visualisation between the Active and Passive groups. Moreover, Constructive students used the personal and others' comments timeline visualisations significantly more than Active students ($p < .001$). Active students also interacted significantly more with the others' comments timeline visualisation than Passive students ($p < .001$). However, Passive students did not use personal timeline visualisation since they did not comment. In Social Space, Constructive students interacted with rating visualisations more than Active students ($p < .001$). However, the Passive group did not interact with any rating visualisation since they had no comment to be rated in the Social Space. In terms of duration of interaction with AVW-Space, ANOVA showed a significant difference between different categories of students. However, the post hoc analysis with the Bonferroni correction showed a significant difference only between Constructive and Passive students in the number of days spent on AVW-Space ($p < .05$).

Table 5-6 Statistical description of interactions performed by students in ICAP categories

Activity	Constructive (N = 185)	Active (N = 141)	Passive (N = 25)	ANOVA
Unique videos	7.50 (1.49)	6.80 (2.44)	4.04 (3.24)	F = 31.59, p < .001
Times tutorial videos watched	6.31 (2.86)	5.89 (3.09)	3.16 (2.62)	F = 12.72, p < .001
Times example videos watched	4.7 (2.40)	4.40 (2.94)	1.92 (2.64)	F = 12.13, p < .001
Quality nudge	14.77 (9.20)	8.53 (5.48)	2.00 (1.50)	F = 47.93, p < .001
Reminder nudge	13.51 (5.64)	10.68 (5.81)	6.80 (6.50)	F = 20.05, p < .001
Comments	19.17 (12.17)	9.80 (6.63)	0	F = 64.25, p < .001
Tutorial low-quality comments	3.98 (4.18)	4.14 (3.18)	0	F = 14.38, p < .001
Tutorial high-quality comments	6.47 (4.00)	1.24 (.95)	0	F = 146.55, p < .001
Ratings	32.18 (68.55)	16.05 (20.24)	1.96 (3.62)	F = 6.29, p < .01
Progress visualisation	9.73 (8.76)	7.24 (6.84)	6.00 (7.56)	F = 5.17, p < .01
Others' timeline visualisation	33.06 (22.47)	18.99 (13.67)	4.72 (5.69)	F = 39.69, p < .001
Personal timeline visualisation	1.73 (2.32)	.58 (1.11)	0	F = 20.76, p < .001
Nudge visualisation	1.00 (1.64)	.93 (1.85)	1.8 (3.09)	F = 2.33, p = .099
Rating visualisation	2.25 (4.78)	.63 (2.08)	0	F = 9.58, p < .001
Days	4.66 (2.98)	4.02 (2.66)	3.16 (3.55)	F = 4.03, p < .05

I used the ENA1.7.0 Webtool to compare the cooccurrences of students' interactions in different ICAP categories in the 2021 study. I used logs of interactions with AVW-Space for ENA. The units of analysis were all lines of data associated with a single ICAP category, subsetted by student_id. I defined the conversation as the lines of logs related to a student_id, subsetted by their sessions with stanza = 2, meaning the aggregation of each line of logs plus the one previous line within a given conversation. I chose this stanza to capture consequent events. The ENA model included the following codes presented in Table 5-7.

Table 5-7 Description of codes derived from event logs of the 2021 studies

Codes	Description	Example in the logs
Personal Space events		
Video_ page	Loading a video page	Pageload: Video Watch
Nudge	Receiving a nudge	Nudge
Comment	Making a comment	Comment created
Others_timeline	Hovering over others comments timeline visualisation	Interaction with visualisation_name= Others_comments_timeline
Personal_timeline	Hovering over personal comments timeline visualisation	Interaction with visualisation_name= Personal_comments_timeline
Progress_vis	Hovering over progress report visualisations	Interaction with visualisation_name= progress_report_visualisation
Nudge_vis	Clicking on nudges visualisations	Interaction with visualisation_name= nudge_visualisation
Social Space events		
Review_ page	Loading the rating page for a video	Video Review
Rating	Rating a comment	Rating_id=6, comment_id=1190
Rating_vis	Hovering over rating visualisations	Interaction with visualisation_name= rating_visualisation

The generated ENA model had co-registration correlations of .97 (Pearson) and .97 (Spearman) for the first dimension and co-registration correlations of .97 (Pearson) and .98 (Spearman) for the second. These measures indicate strong goodness of fit between the network visualisation and the original model. Figure 5-4 shows the network for all students. The strongest connections are the interaction between *Nudge*, *Others_timeline* and *Comment*. Also, the connections for *Review_ page – Rating* and *Progress_vis – Video_ page* are strong for all students. The centroid of the network for Constructive is in the top half of the projected space, where *Personal_timeline*, O*thers_timeline* and *Rating_vis* are positioned. However, the centroid of the network for Passive students is in the bottom left quartile of the space and is close to the *Review_ page*, *Video_ page* and *Progress_vis.* The network centroid of Active students is also in the bottom half of the projected space, but it is close to the centre.

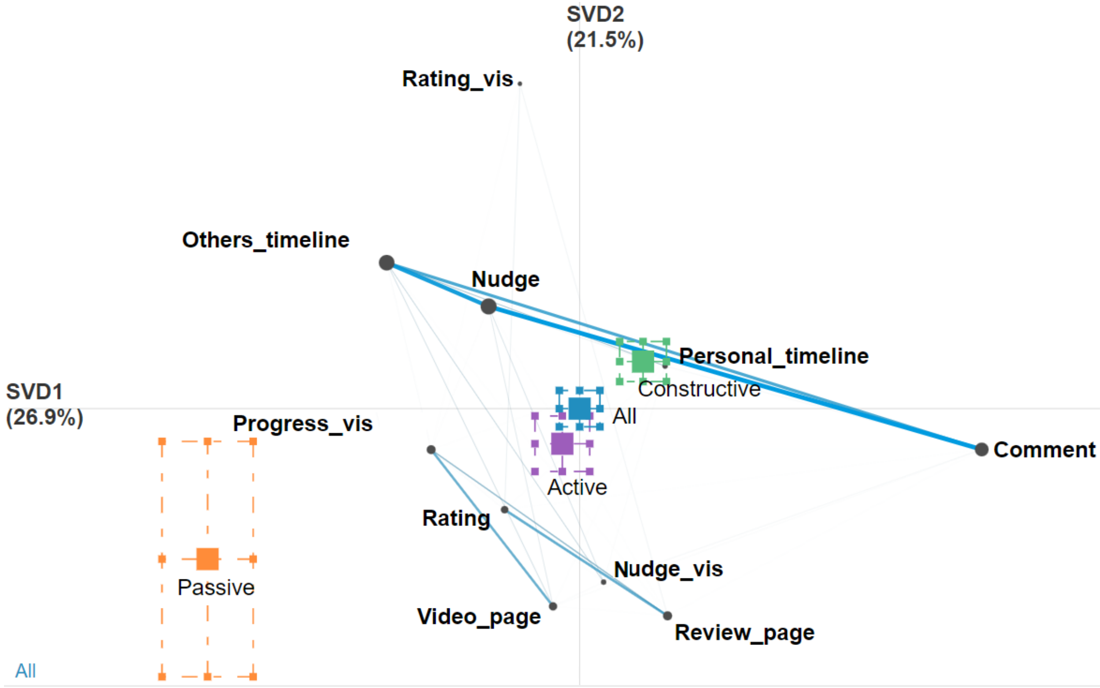

Figure 5-4 ENA for all students and the centroids for ICAP categories

To test the differences between Constructive and Active (Figure 5-5), I applied a two-sample t-test on the location of points in the projected ENA space. There were statistically significant differences between Constructive (mean = .31, SD = .92, N = 185) and Active (mean = -.11, SD = 1.09, N = 141) along the X-axis (t = -4.43, p < .001, Cohen's d = .50) and along the Y-axis (t = 4.74, p < .001, Cohen's d = .54). Figure 5-5 shows that the connections of *Personal_timeline* and *Others_timeline* with *Comment* and the connection between *Comment* and *Nudge* are stronger for Constructive students than Active students. In contrast, the connections for *Progress_vis – Video_page, Nudge – Nudge_vis*, *Nudge_vis – Others_timeline* and *Nudge_vis – Comment* are stronger for Active students. This could indicate that nudge visualisations were more beneficial for commenting in the Active group. In contrast, Constructive students commented more when receiving nudges directly and interacting with timeline visualisations. In addition, Constructive students have a strong connection for *Rating_vis – Rating*, while Active students had a stronger connection for *Review_page – Rating*. Thus, Active students did not interact with rating visualisations in reviewing phase as frequently as Constructive students.

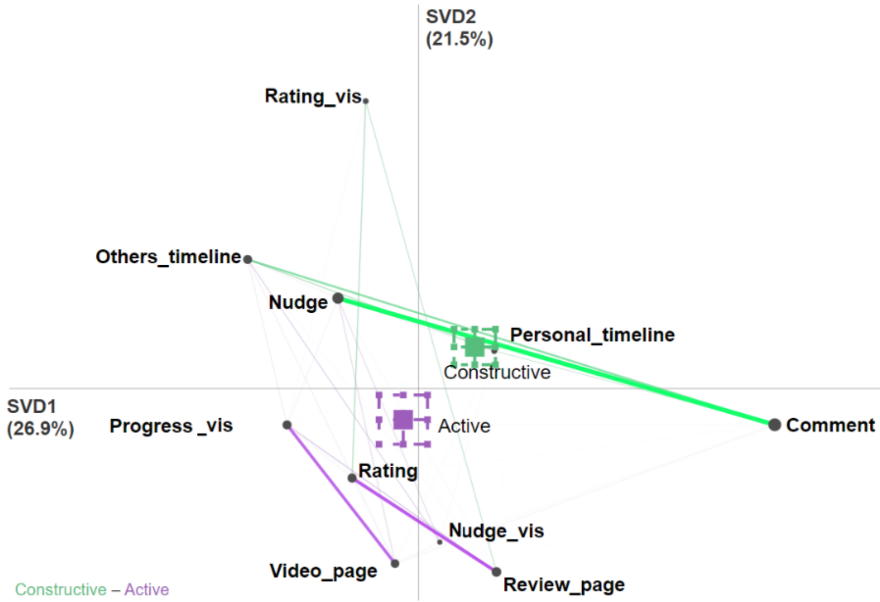

Figure 5-5 Difference network for Active and Constructive students

I also investigated the differences between the Active and Passive groups (Figure 5-6). The t-test showed statistically significant difference between Passive (mean = -2.47, SD = .73, N = 25) and Active along X-axis; (t = 13.61, p < .001, Cohen's d = 2.25), but the difference along the Y-axis was not significantly different (t = -1.96, p = .06, Cohen's d = .61). Figure 5-6 shows that all connections are stronger for Active students except *Progress vis – Video_ page*, *Nudge – Nudge_vis* and *Nudge – Others_timeline*, which are stronger for Passive students. Although Passive students did not make any comments, they hovered over the nudge visualisations to review what nudge was received. They also hovered over the others' comments timeline when getting nudges, but these interactions with visualisations did not result in commenting for the Passive group.

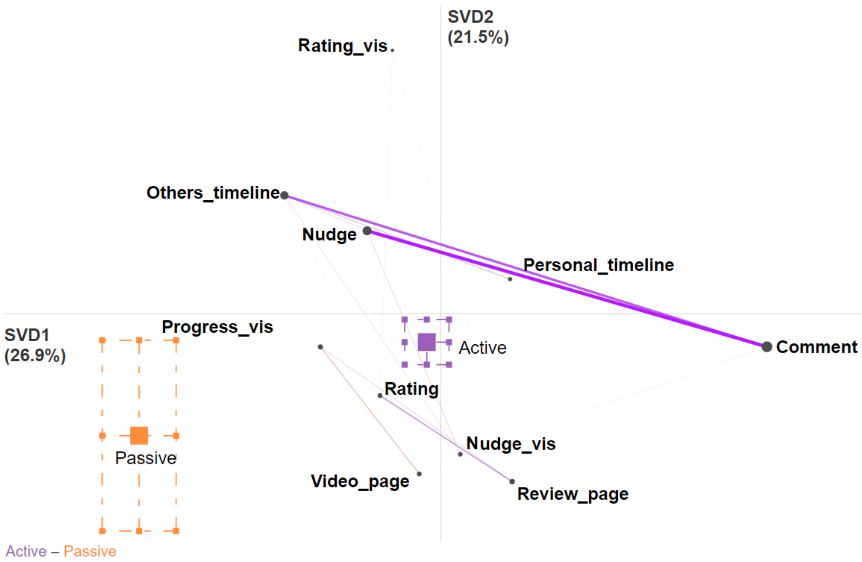

Figure 5-6 Difference network for Active and Passive students

A two-sample t-test assuming unequal variance showed a statistically significant difference between Constructive and Passive students along the X-axis (t = -17.37, p < .001, Cohen's d = 2.79) and Y-axis (t = 3.41, p < .001, Cohen's d = 1.22). The difference network for Constructive-Passive is similar to the difference network for Active-Passive. The analyses provided in Table 5-6 and ENA results confirm hypothesis H5.3.

5.3.2 RQ5.2: Effects of Visualisations on Learning

Table 5-8 shows the pre- and post-study conceptual knowledge scores in the 2020 and 2021 studies for those who completed both surveys and interacted with AVW-Space. The independent t-test between the two studies showed no significant difference in their CK1 (t = 1.61, p = .10). After applying ANCOVA on CK2 between two studies and controlling CK1, I found no significant difference in CK2 (F = .6, p = .80).

Table 5-8 Statistical description of conceptual knowledge of students in two studies

Study	CK1	CK2
2020 (N=147)	13.14 (5.08)	13.55 (5.71)
2021 (N=277)	14.18 (6.05)	13.53 (6.47)

The previous chapter showed significantly higher CK2 for Constructive students than Active and Passive students. Thus, I expected to see a significant difference between the learning of the Constructive and other categories in the 2021 study (Table 5-9). There were no significant differences on CK1 between ICAP categories. However, there was a significant difference on CK2 between categories. Post-hoc analysis using Bonferroni correction showed significantly higher CK2 for Constructive students compared to Active students. There was no significant difference on CK2 between Passive and Active students; however, only six Passive students completed both surveys. This result is consistent with previous studies, where learning outcome was higher for students who made more high-quality comments. The ANCOVA on CK2 scores of Constructive students in the 2020 (15.92 ± 5.85) and 2021 studies showed no significant difference. ($F = 1.28$, $p = .25$).

Table 5-9 Statistical description of conceptual knowledge of students in ICAP categories

CK	Constructive (N = 153)	Active (N = 118)	Passive (N = 6)	Significance
CK1	13.79 (6.21)	14.49 (5.62)	18.33 (6.15)	$F = 1.89$, $p = .15$
CK2	14.76 (6.45)	12.10 (6.30)	10.16 (2.48)	$F = 7.1$, $p < .001$

The structural equation model for learning in the 2020 study indicated the direct effect of Quality nudges on the number of high-quality comments, affecting CK2. However, that model does not fit the 2021 data. Therefore, I developed a model for the 2021 study illustrated in Figure 5-7 to address hypothesis H5.4. The chi-square test (14.01) for this model (DF = 9, 19 estimated parameters) shows that the model's predictions were not statistically significantly different from the data ($p = .12$). The Comparative Fit Index (CFI) was .991, and the Root Mean Square Error of Approximation (RMSEA) was .04. Hence, the model is acceptable: CFI is greater than .9, and RMSEA is less than .06 (Hu & Bentler, 1999). The model indicates that a higher number of high-quality comments causes a higher CK2 score ($p < .01$). Moreover, the number of interactions with rating visualisations positively affects CK2 ($p < .001$). Other links are all significant at $p < .001$ except progress visualisation → Video ($p < .05$) and progress visualisation → Personal Space visualisations ($p < .05$). The covariances with e8 show that a student who interacts with one visualisation is likely to interact with other visualisations. The model illustrates that the number of videos watched, received nudges, and interactions with visualisations affect the number of high-quality comments and, consequently, CK2. Thus, Hypothesis 5.4 is confirmed.

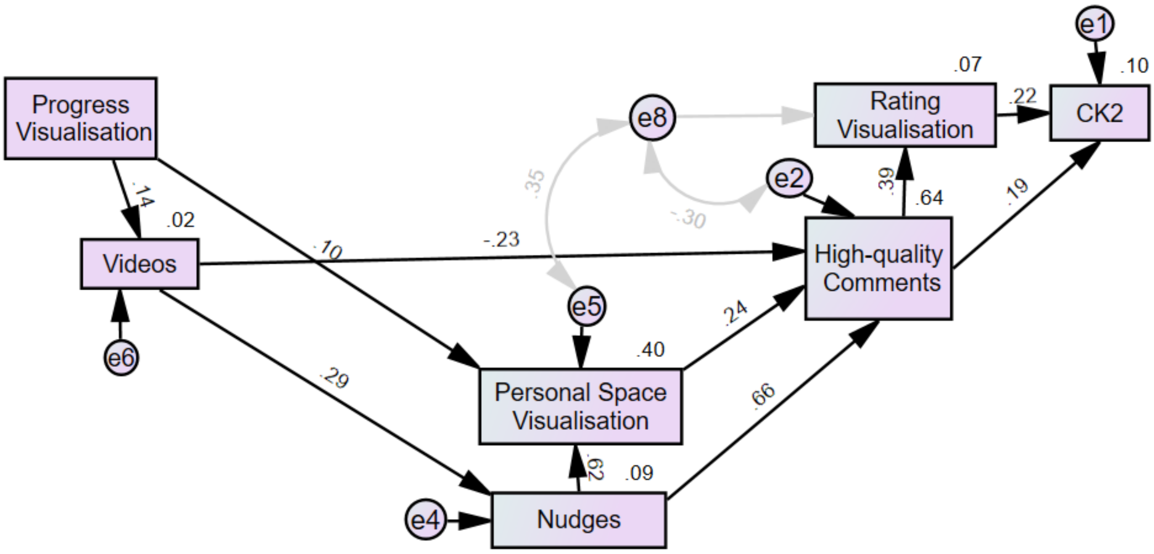

Figure 5-7 Path diagram for testing hypothesis H5.4: visualisations increase the number of high-quality comments and learning

5.3.3 RQ5.3: Subjective Opinions on Visualisations

I compared the responses to the TLX-NASA questionnaire from the 2020 and 2021 studies. There was no significant difference in the scores for the cognitive load of commenting tasks between the 2020 and 2021 studies. However, the 2021 participants found rating significantly less mentally demanding (7.27 ± 4.10) and effortful (6.77 ± 4.90) compared to the 2020 study (8.02 ± 3.84 and 7.77 ± 5.07, respectively); ($t = 4.02$, $p < .001$ and $t = 3.17$, $p < .01$, respectively). I also compared the perceived usefulness of AVW-Space in the two studies. As shown in Table 5-10, the t-test showed that responses to the TAM statements 1, 2, 4-7 from the 2021 study were significantly higher than those of the 2020 study. However, the mean differences were small and less than one score. In addition, ANOVA on the TAM usefulness responses, with the year of study and ICAP category as two fixed factors, revealed that the interaction of these factors (year*ICAP category) has no significant effect on the responses to the TAM questions.

Table 5-10 TAM usefulness scores in the 2020 and 2021 studies

TAM Statement	2020 (N = 147)	2021 (N=277)	Significance
1. I think I would like to use AVW-Space frequently.	4.04 (1.565	4.43 (1.45)	t = 2.60, $p < .05$
2. I would recommend AVW-Space to my friends.	3.78 (1.44)	4.45 (1.53)	t = 4.35, $p < .001$
3. Using AVW-Space would enable me to improve my soft skills quickly.	3.16 (1.38)	4.40 (1.36)	t = 1.74, p = .07
4. Using AVW-Space would improve my performance, considering the development of soft skills.	3.0 (1.31)	2.34 (1.36)	t = 2.51, $p < .05$
5. Using AVW-Space would enhance my effectiveness when developing soft skills.	3.00 (1.21)	3.46 (1.34)	t = 3.52, $p < .001$
6. I would find AVW-Space useful in my studies/job.	3.26 (1.44)	3.72 (1.45)	t = 3.10, $p < .01$
7. I would find AVW-Space easy to do what I want it to do.	3.34 (1.41)	3.81 (1.46)	t = 3.22, $p < .01$
8. My interaction with AVW-Space would be clear and understandable.	3.24 (1.36)	3.49 (1.36)	t = 1.82, p = .07
9. I would find AVW-Space easy to use.	2.97 (1.41)	3.20 (1.41)	t = 1.58, p = .11
10. If I am provided the opportunity, I would continue to use AVW-Space for informal learning.	3.91 (1.59)	4.18 (1.62)	t = 1.61, p = .10

I collected the students' feedback on visualisations from Survey 2 in the 2021 study. The progress visualisation received the most positive feedback (90.97%) among all visualisations. 252 students reported that the progress visualisations increased their motivation, clarified tasks to complete and facilitated learning organisation and time management. For example, a student said: "*It is a great way to keep track of progress and motivated me to get it finished. I was about to leave the last two videos until I saw how close I was to finishing, so I kept going.*".

75% of feedback on the timeline visualisations was positive, stating that the timeline visualisations helped them recognise important parts of the video and inspired them to make comments on different topics (e.g., *"It initially helped me to grasp an idea of what kind of comments were being looked for. It also encouraged me to try pick up on points in areas of the video which had less comments made.").* The timelines also helped students to compare their progress with other students (e.g., *"... I could see if the type of comments I was writing were on the right track and if I was making enough comments throughout the video.").* However, some students found the timeline visualisations cluttered and suggested adding filtering functionality for the comments presented in the timeline visualisation. Some students also pointed that they did not like the others' comments timeline visualisation since they wanted to form their own thoughts.

70% of feedback on nudges visualisations and comment quality indicators was positive. Students noted using these visualisations as feedback to improve commenting and stay on track. For instance, a student said: "*This tool can help you notice a pattern in the nature of hints you are getting, to identify whether there is something you consistently fail to recognise, or something you always comment on.*". However, students who criticised the quality indicators suggested more explanation on their meaning and criteria.

Lastly, 69% of feedback on rating visualisations were positive since they helped students to reflect whether their comments were useful to their peers and encouraged students to make better comments. Some examples of positive feedback on rating visualisations are: " *... If I had an unpopular opinion I could see this from the ratings and investigate reasons why.*" and *"To motivate people to write good comments and also so they feel good about the work they've done"*. On the other hand, 14% of students found the rating visualisations unhelpful since they disliked being judged and received a limited number of ratings.

I also investigated the feedback on nudges for the 2020 and 2021 studies. In the 2020 study, 69% of feedback on nudges were positive, but the percentage of positive feedback on the nudges increased to 74.72% in the 2021 study. This could be the result of showing the record of nudges in the comment list visualisation and helping students review their received nudges. The students' feedback, as well as the comparison of TAM and TLX-NASA scores in the two studies, provide evidence supporting hypothesis H5.5.

5.3.4 RQ5.4: Effectiveness of Visualisations for EFL Students

In order to investigate the effectiveness of visualisations on the engagement of EFL/Native students, I first investigated the distribution of EFL and Native students in the 2020 and 2021 studies. There were 29 EFL students in the 2020 study, and 42 EFL students in the 2021 study. In other words, 9.9% of 2020 participants and 11.9% of 2021 participants were EFL students. The chi-square test of homogeneity between being EFL/Native and the studies showed no significant difference in the distribution of EFL/Native between the two studies (Chi-square = .74, p = .39; Phi = .34, p = .39). Thus, I could compare the distribution of ICAP categories for EFL and Native students in both studies to assess the effects of visualisations (Table 5-11). For Native students, the Chi-square test of homogeneity between the studies and ICAP categories showed a significant difference (Chi-square = 34.61, $p < .001$), with a significant effect size of Phi = .24, $p < .001$. For EFL students, chi-square test between the studies and ICAP categories also showed a significant difference (Chi-square = 14.26, $p < .001$) with an effect size of Phi = .44 and $p < .001$. Post hoc analysis using z-test with a Bonferroni correction revealed a significant increase in the proportion of the Constructive group and a significant decrease in the proportion of Passive students for both EFL and Native students in the 2021 study compared to the 2020 study ($p < .05$). Moreover, the chi-square test between being EFL/Native and the ICAP categories for the 2021 study showed no significant difference (Chi=2.31, $p = .34$, Phi = .07), unlike the previous studies where most EFL students were in the Passive group. Hence, the visualisations were effective in enhancing the engagement of EFL students. I investigated EFL and Native students' interactions with visualisations. Independent t-test showed no significant difference in the EFL/Native students' number of interactions with visualisations except for the rating visualisations ($t = 2.31$, $p < .05$). Native students interacted with rating visualisations (5.01 ± 14.04) significantly more than EFL students (2 ± 6.64).

Table 5-11 Distribution of ICAP categories in EFL and Native students for the 2020/2021 studies

	ICAP categories	2020	2021
EFL	Passive	15 (51.7%)	5 (11.9%)
	Active	9 (33.3%)	18 (42.9%)
	Constructive	5 (17.2%)	19 (45.2%)
Native	Passive	60 (22.6%)	20 (6.5 %)
	Active	105 (39.6%)	122 (39.6%)
	Constructive	100 (37.7%)	166 (53.9%)

I compared the effectiveness of visualisations on the learning of EFL and Native students who completed both Surveys 1 and 2 in the 2021 study (Table 5-12). There was no significant difference in the CK1 of EFL and Native students ($F = 3.32$, $p = .07$), similar to the previous study. However, ANCOVA on CK2, between EFL and Native students, with controlling CK showed that Native students had significantly higher CK2 than EFL students ($F = 4.25$, $p = .04$). However, the lower scores in CK2 for EFL students could be due to the small population size of EFL students and their language difficulties in answering the conceptual knowledge questions. The results presented in Table 5-11 showed visualisations improved the engagement of EFL and Native students, but Table 5-12 revealed that the learning outcome for Native students was significantly more than EFL students, after receiving the visualisations. Thus, hypothesis H5.6 is addressed.

Table 5-12 Conceptual knowledge of EFL and Native students before and after the 2021 study

	CK1	Ck2
Native (N = 241)	14.41 (6.13)	13.83 (6.36)
EFL (N = 35)	12.42(5.26)	11.31(6.88)

5.4 Conclusions

In this chapter, I proposed the integration of various visual learning analytics into VBL to increase engagement and learning. These visualisations are intended to assist students in monitoring and managing their learning progress in VBL. I investigated the effectiveness of these visualisations by conducting a study with the new version of AVW-Space. The study revealed that the visualisations effectively enhance engagement in VBL: participants watched more videos, wrote more comments, used more diverse rating options and spent more days on AVW-Space when receiving the visualisations compared to the previous study. Furthermore, the visualisations increased the percentage of students who showed constructive engagement and made high-quality comments. The visualisations also increased the number of Constructive EFL students. The participants who made more high-quality comments also interacted more with the visualisations and significantly increased their conceptual knowledge after using AVW-Space. Participants who received the visualisations also perceived rating less effortful and mentally demanding and found visualisations useful. The increase in engagement and learning outcome in the study with the visualisations is consistent with the literature findings on the effectiveness of visual learning analytics in other computer-assisted teaching environments. In addition, the students' feedback shows that the visualisations boosted students' motivation and self-reflection, as litrature suggested.

Although most participants found VLA motivating and beneficial in learning organisation, I discovered some challenges in the integrated visualisations. Firstly, some students were unable to interpret the quality indicators of comments. As literature also notifies the importance of the explainability of visual learning analytics, future versions need to clarify the quality criteria to the students. I also noticed that some students found visualisations of others' comments or visualisations of peer's ratings stressful and interfering with their knowledge construction, while some other students found these visualisations very helpful for reflecting on their learning progress. Therefore, the future version of AVW-Space could offer a visibility option for visualisations that triggers social comparison. Future work could also involve further investigation of students' learning strategies and tailoring visualisations visibility to the students' behaviours. Another challenge in visualisations is that teachers must manually select comments shown in the others' comments timeline, which could be tedious for a large class. Therefore, potential solutions for automating this task could be explored in future research.

The main challenge in studying the effectiveness of visualisations is how to measure the interactions with visualisations. Since most interactions are in the form of hovering, it is difficult to identify which interactions were intentional and focused. One approach for investigating the interactions more precisely is analysing eye-gazing, but this approach is time-consuming and unpractical for a large class. Another limitation in this study is that the population were first-year engineering students. Therefore, the effectiveness of the visualisations should be investigated for students in other courses. The next chapter will address this limitation by investigating the effectiveness of the visualisations for another course.

6 Generalisability of Visualisations and Nudges

To investigate the generalisability of nudges and visualisations, I conducted a new study using the enhanced version of AVW-Space in a different domain. The domain chosen for this purpose was communication skills for face-to-face meetings in software engineering projects. Communication is a crucial soft skill in software engineering since it promotes information sharing with stakeholders (Prenner et al., 2018). However, teaching this skill is time-consuming, involves hands-on exercises and regular feedback from teachers (Anthony & Garner, 2016; Galster et al., 2018; Sedelmaier & Landes, 2018). AVW-Space can offer a video-based approach for teaching communication skills by encouraging self-reflective learning. A recent study investigated the use of AVW-Space to train face-to-face meeting communication skills for software engineering students (Musa et al., 2021). However, the version of AVW-Space in that study did not include nudges or visualisations. Therefore, this chapter presents the generalisation of nudges and visualisations for training communication skills, and investigates the students' behaviour in the context of communication skills to address the last two research questions of my PhD research:

RQ6. How can the quality assessment models, nudges and visual learning analytics be generalised to other soft skills? The quality assessment schemes proposed in Chapter 3 are domain-independent. In addition, the classifiers for quality assessment of comments use the domain-specific ratio of comments and some LIWC features that are also independent of the domain. Therefore, I first hypothesised that the quality assessment models proposed in Chapter 3 could be generalised by calculating the domain-specific ratio of comments based on the domain vocabulary (hypothesis H6.1). In other words, I hypothesised that the quality assessment models would perform well for comments on communication skills if I use the models trained on presentation skills, but calculate the domain-specific features according to the domain vocabulary of communication skills.

RQ7. Do nudges and visual learning analytics increase engagement and learning in other soft skills? I narrow this question to the communication skills domain. After customising nudges and visualisations for communication skills, I conducted a new study to address the following subquestions:

RQ7.1. Do nudges and visual learning analytics increase students' engagement when using AVW-Space for training communication skills? Previous studies with AVW-Space showed that nudges and visual learning analytics increased engagement measured by the number of watched videos, comments or ratings. I expected to see increased number of videos, comments, high-quality comments, and higher diversity of rating options and aspects, after providing the customised nudges and visualisations (hypothesis H7.1). In addition, I anticipated more students with constructive commenting behaviour resulting from nudges and visualisations in communication skills (hypothesis H7.2). I hypothesised that there would be a difference in engagement for ICAP categories in the communication skills context, as measured by the number of high-quality comments and interaction with visualisations (Hypothesis H7.3).

RQ7.2. Do nudges and visual learning analytics increase learning of communication skills via AVW-Space? The previous study on communication skills without nudges and visualisations showed a significant increase in the students' knowledge (Musa et al., 2021). Hence, I expected an increase in students' knowledge after training with generalised nudges and visualisations (hypothesis H7.4). I also anticipated an increase in knowledge of Constructive students who made high-quality comments (hypothesis H7.5). I expected a causal relationship between the visual learning analytics and nudges provided for communication skills and students' knowledge. Hence, I hypothesised that visual learning analytics and nudges would increase the number of high-quality comments and increase learning (Hypothesis H7.6).

RQ7.3. Do students behave differently in the context of presentation skills and communication skills? I expected to see differences in students' interactions with visualisations and nudges between two contexts of communication and presentation skills (hypothesis H7.7).

Section 6.1 describes the initial study on communication skills without nudges or visual learning analytics. Then, Section 6.2 explains how nudges can be generalised and customised to the communication skills domain by leveraging the data collected from the initial study (RQ6). Then, Section 6.3 presents the experimental design of the study with customised nudges and visualisations, followed by a comparison of engagement and learning in the two studies. The results of statistical analyses are presented in Section 6.4 to address RQ7.1 – RQ7.3. Finally, Section 6.5 summarises the results, discusses the challenges and limitations in generalising nudges and visualisations, and suggest opportunities for future improvement.

6.1 Initial Study on Communication Skills without Nudges and Visualisations

The first study for training face-to-face meeting communication skills using AVW-Space was conducted in a second-year Software Engineering project-based course (SENG202) at the University of Canterbury in 2020. This study was administrated by Jaafaru Musa, as a part of his PhD research. The Human Ethics Committee of the University of Canterbury approved this study with the reference “HEC 2020/30/LR-PS”. I refer to this study as the “SENG2020 study”. The course runs over one semester, where students work in teams of four to six, and have weekly face-to-face meetings. There were 56 students enrolled in SENG202 in 2020. The students who participated in the study and watched videos received 1% credit of the final grade.

At the start of the study, Survey 1 was administrated, which included questions on demographic, training, experience with face-to-face meetings and working in non-academic software development teams; a question relating to participant’s conceptual knowledge of face-to-face meetings communication skills; how frequently they watch YouTube in general and for learning; followed by a self-reported face-to-face meeting communication scale developed for this study. In the question for conceptual knowledge, the participants had one minute to list all concepts they knew about the face-to-face meetings communication skills. After Survey 1, the participants watched and commented on ten carefully selected short videos on effective meeting communication. As listed in Table 6-1, there were six tutorial videos on communication skills and four example videos of real meetings. The aspects defined for tutorial videos were the same as those for presentation skills: “*I didn’t realise I wasn’t doing it*”, “*I am rather good at this*”, “*I did/saw this in the past*”, and “*I like this point*”. However, the aspects for the example videos were “*Verbal communication*”, “*Giving feedback*”, “*Active listening*”, “*Meeting contribution*”, and “*Receiving feedback*”. The Personal Space in the SENG2020 study did not provide any nudges or visualisations.

Table 6-1 Description of videos used for face-to-face meeting communication skills on AVW-Space

Video	Title	Length (s)	YouTube ID
Tutorials			
1	The 7 Cs of Communication	166	sYBw9-8eCuM
2	Body Language	165	AqixzdpJL4U
3	Improve Your Listening Skills with Active Listening	160	t2z9mdX1j4A
4	Giving feedback	106	Id_uG8Djdsc
5	How to effectively contribute to team meetings	245	cKh75Po5Qsc
6	How Google builds the perfect team	143	v2PaZ8Nl2T4
Examples			
1	Examples of Good Meeting Communications Skills	110	czpBKC9Plh4
2	Bad Stand-up	322	zrmcl-pjmoc
3	The Daily Stand-up Meeting	154	VjNxQ-a-x2M
4	EXAMPLE 4 – How NOT to run a meeting	147	F1qstYxrqn8

The SENG2020 study had an additional phase (phase 3) where each team recorded one of their weekly meetings and commented on the recording of their own meeting, and subsequently rated comments written by their teammates. However, the third phase is out of the scope of this PhD research. Finally, Survey 2 was administrated, consisting of the same questions on participants' conceptual knowledge of communication skills and the self-reported scale. Survey 2 also had three other questionnaires: CAP perceived learning gain scale (Rovai et al., 2009); NASA-TLX (Hart, 2006) cognitive load scale; Technology Acceptance Model (TAM) (Davis, 1989) scale to capture students' overall perception of AVW-Space; and questions on usability of the AVW-Space. In the SENG2020 study, 47 students completed Survey 1, and thirty of them completed both Surveys 1 and 2.

6.2 Generalising Nudges and Visualisations

In this section, I first investigate how quality assessment models developed for presentation skills perform in the communication skills context. After generalising the quality assessment models, I present the enhancements of AVW-Space to make nudges and visualisations customisable for different domains, including communication skills.

6.2.1 RQ6: Generalising the Quality Assessment Models

After exploring comments made in the SENG2020 study, I discovered that the quality scheme developed for presentation skills could be applied to comments made on face-to-face communication videos. Therefore, I manually labelled the comments from the SENG2020 study. Table 6-2 shows the frequency of quality categories in comments from the SENG2020 study. Similar to the previous studies on presentation skills, category 2 (repeating video content) was the most frequent category.

Table 6-2 Distribution of comments in quality categories for tutorial and example videos

Tutorial comment categories	1	2	3	4	5
Count	1	70	18	51	19
Percentage	.6%	44.02%	11.32%	32.07%	11.94%
Example comment categories	1	2	3		
Count	5	220	67		
Percentage	1.90%	75.34%	22.94%		

In order to evaluate the effectiveness of quality assessment models, I first needed to extract the domain vocabulary for communication skills. Therefore, I generated the corpus from the transcripts of tutorial videos (N = 296 segmented text), answers to conceptual knowledge questions (N = 49) and comments on the tutorial (N = 159) and example (N = 292) videos. After lowercasing texts, removing punctuations and stop words (e.g., “am”, “is”, “to”), the tokens were extracted and lemmatised. Next, words and bigram phrases that appeared more than twice in the corpus were extracted automatically using collocation statistics (Mikolov et al., 2013) implemented in the Phrases module of the Genism library[7]. For each word, I extracted the most relevant and similar words using GloVe (Global Vectors for Word Representation)[8] (Pennington et al., 2014) word vectors pre-trained on Wikipedia 2014[9] and Gigaword[10]. Finally, 225 words and phrases were extracted from the texts, along with 225 synonyms defined for them.

Next, three expert coders, including the course coordinator, verified whether the extracted words should be in the domain vocabulary. The pairwise Cohen’s Kappa test revealed moderate (.55), substantial (.61) and nearly perfect (.91) agreement between the coders (Landis & Koch, 1977). Similarly, the Fleiss’ kappa showed substantial inter-coder agreement for the extracted words (κ = .69) (Landis & Koch, 1977). However, Krippendorff’s alpha coefficient (α = .31) was lower than the minimum acceptable value (α > .66) for inter-coder agreement (“Krippendorff’s Alpha,” 2010). Therefore, a meeting was organised for the three coders to review and discuss their codes and achieve agreement. After the meeting, eleven words were discarded, and ten new words were added to the domain vocabulary.

Next, the domain-specific ratio was computed for all comments made for tutorials and example videos in the SENG2020 study. Then, the calculated domain-specific ratios with extracted LIWC features of comments from the SENG2020 study were used to test the quality assessment models trained on presentation skills. The quality assessment models performed well with weighted F1-scores of .74/.98 for comments made on tutorial/example videos. The confusion matrices of quality assessment models are presented in Figure 6-1. The confusion matrices indicate the good performance of the quality assessment models in the communication skills context. Therefore, hypothesis H6.1 is confirmed.

---

[7] https://radimrehurek.com/gensim/models/phrases.html
[8] https://nlp.stanford.edu/projects/glove/
[9] https://dumps.wikimedia.org/enwiki/20140102
[10] https://catalog.ldc.upenn.edu/LDC2011T07

$$\begin{array}{rc} \mathit{classified\ as:} & \begin{array}{ccc} 1 & 2+3 & 4+5 \end{array} \\ \begin{array}{r} \mathit{actual} = 1 \\ \mathit{actual} = 2+3 \\ \mathit{actual} = 4+5 \end{array} & \begin{pmatrix} 0 & 1 & 0 \\ 1 & 74 & 13 \\ 0 & 25 & 45 \end{pmatrix} \end{array} \qquad \begin{array}{rc} \mathit{classified\ as:} & \begin{array}{cc} 1 & 2+3 \end{array} \\ \begin{array}{r} \mathit{actual} = 1 \\ \mathit{actual} = 2+3 \end{array} & \begin{pmatrix} 2 & 3 \\ 0 & 287 \end{pmatrix} \end{array}$$

Figure 6-1 Confusion matrix for comments made on the tutorial (left) and example (right) videos in communication skills

After confirming hypothesis H6.1, the AVW-Space nudge engine was enhanced by generalising its quality prediction module. Previously, the quality prediction module calculated the domain-specific ratios by using a text file of domain vocabulary and phrases for presentation skills. In the generalised version of the nudge engine, the quality prediction module can calculate the domain-specific ratio of comments using any text file of domain vocabularies corresponding to the target skill.

6.2.2 Enhancing the Teacher Interface for Customising Nudges and Visualisations

Nudges need to refer to the relevant target skill. Previously, the text for nudges and their examples were hard-coded in AVW-Space for presentation skills. Therefore, the example comments provided in nudges should be relevant to the videos. In addition, comments used in the interactive timeline visualisations were hard coded for only presentation skills. Hence, I enhanced the teacher interface for customising the nudges to the instructional domain.

Figure 6-2 shows the teacher interface for communication skills before and after enhancing the teacher interface and making nudges and visualisations customisable. In the enhanced teacher interface, I added more options for enabling/disabling different types of support (visualisations, nudges, and types of nudges). The enhanced interface also allows the teacher to upload a file of example comments to be displayed in the interactive timeline visualisation. For each video, there are three nudges (*No Comment*, *Elaborate More* and *Critical Thinking*), which require curated example comments. If the teacher enables commenting and Quality nudges, these examples need to be provided for each video. In addition, the *Aspect Under-utilised Nudge* provides an explanation of the aspect that the student has not used enough. Again, if the teacher has enabled the nudges for aspects, the explanation fields for aspects need to be filled.

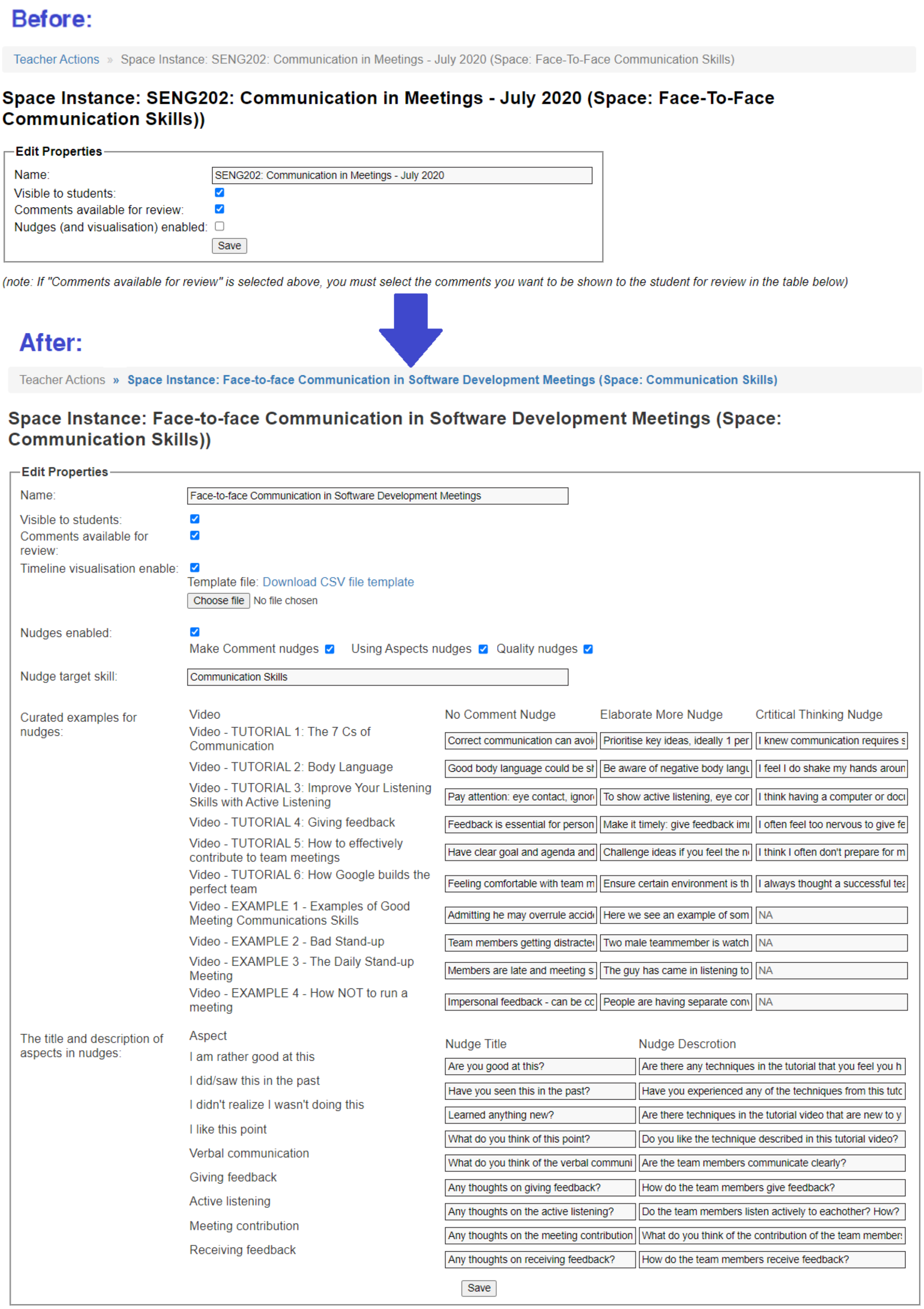

Figure 6-2 Teacher interface before and after making nudges and visualisations customisable for different domains

For using AVW-Space on training communication skills, the example comments for nudges were defined using comments made in study SENG2020 as shown in Figure 6-2. Next, the Personal and Social Space were evaluated by three domain experts, who found visualisations and nudges functioning similarly to the visualisations and nudges for presentation skills. Figure 6-3 shows the Personal Space for communication skills with customised nudges and visualisations.

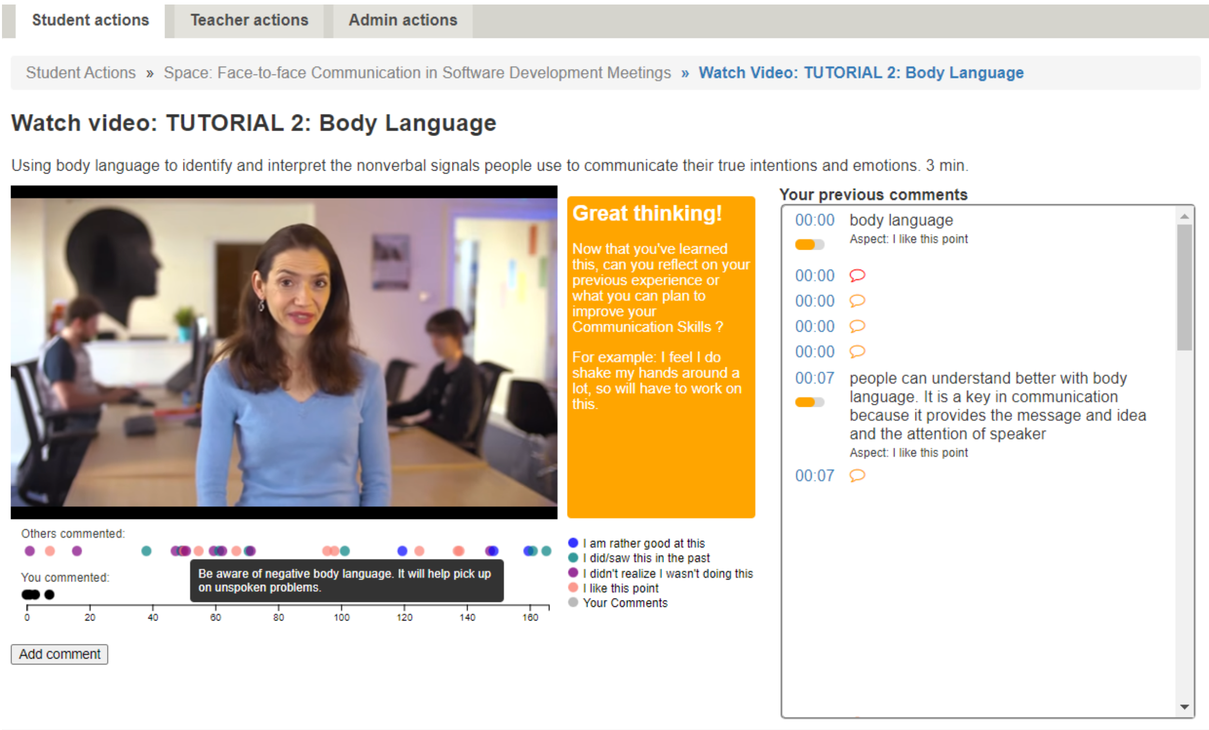

Figure 6-3 Personal Space for communication skills space, with customised nudges and visualisations

6.3 Experimental Design

After customising the nudges and visualisations for communication skills, a study was conducted in the same course (SENG202) in 2021 to investigate their effectiveness. This study was also administrated by Jaafaru Musa, as a part of his PhD research. I refer to this study as the “SENG2021 study”. In this study, students were invited to use the enhanced version of AVW-Space, which included customised nudges and visualisations for face-to-face meeting communication skills. The participants first completed Survey 1, which was identical to the SENG2020 study. Next, students were instructed to watch videos, make comments in the first phase, and rate their peers’ comments in the second phase of the study. The first and second phases of the SENG2021 study were identical to the SENG2020 study. However, students were asked to respond to a new survey at the end of the second phase. This new survey included the same questions for participants' knowledge of communication skills, the self-reported communication skills scale, and the CAP perceived learning gain scale. The third phase of the study was identical to the SENG2020 study and was followed by the same final survey, consisting of questions on participants' knowledge of communication skills, the self-reported scale, the CAP perceived learning gain scale, TAM and NASA-TLX questionnaires. The course credit for participating in the study was increased to 5%. The third phase of this study was announced as non-mandatory due to COVID19 restrictions. As the result, only ten students responded to the final questionnaire on the usefulness and cognitive load of AVW-Space.

The next section compares various measurements of engagement and learning of students in SENG2020 and SENG2021 studies. I considered SENG2020 as the control group with no nudges and visualisations, and the SENG2021 study as the experimental group with customised nudges and visualisations.

6.4 Results

In the SENG2021 study, 49 students completed Survey 1 watched the ten videos and made at least ten comments. Thus, there were no Passive students in the SENG2021 study. Only four students did not rate any comments. Forty SENG2021 participants completed Surveys 1 and 2. Table 6-3 shows the demographics of participants who completed Survey 1 and their experience and knowledge in face-to-face communication skills as well as how frequently they watch YouTube. As can be seen, most students were male, native English speakers and aged 18-23 in both studies. The independent t-test showed a significant difference in the participants' meeting experience, non-academic software engineering teamwork experience and using YouTube for learning between the two studies. The participants in the SENG2021 study had significantly higher self-reported scores for meeting experience and the frequency of using YouTube for learning than the SENG2020 study participants. On the other hand, the self-reported score for non-academic experience in software engineering teamwork for the SENG2020 participants was significantly higher than the SENG2021 participants. In order to compare the participants' conceptual knowledge, the students' answers to conceptual knowledge questions were marked automatically, using the vocabulary extracted for communication skills. Then, the marks for conceptual knowledge questions were used as the pre/post-test scores (CK1/CK2). The independent t-test revealed no significant difference in CK1 between the SENG2020 and SENG2021 studies.

Table 6-3 Demographics of students in the SENG2020 and SENG2021 studies (mean and standard deviation)

	SENG2020 (N = 47)	SENG2021 (N = 49)	Significance
Gender	40 males, 8 females	39 male,8 female, 2 other	
Aged 18-23	47	48	
Native English speakers	37	43	
Training	1.34 (.75)	1.34 (.59)	t = .04, p = .963
Meeting experience	1.42 (.92)	2 (1.02)	t = 2.88, p = .005
Non-academic teamwork experience	1.89 (.86)	1.36 (.88)	t = 2.94, p = .004
YouTube	4.38 (.96)	4.53 (.81)	t = .80, p = .421
YouTube for learning	3.23 (1.00)	3.73 (1.05)	t = 2.37, p = .02
CK1	6.76 (4.56)	6.81 (4.66)	t = .05, p = .95

I compared the predicted quality categories for SENG2021 comments to their quality categories defined by a human coder. For comments on tutorial videos, the quality assessment model performed well with F1-score of .79. There were no comments in category 1 (off-topic or short affirmations or negations) on example videos, and the model predicted them all correctly. Hence, the generalised quality assessment models performed well on comments from the SENG2021 study.

6.4.1 RQ7.1: Effects of Nudges and Visualisations on Engagement

Table 6-4 compares students' activities and the duration of interactions with AVW-Space in the SENG2020 and SENG2021 studies. The independent t-test showed that the number of days and sessions spent on AVW-Space were not significantly different between the two studies. However, the SENG2021 participants watched significantly more unique videos and made significantly more comments than participants in the SENG2020 study. In addition, students made significantly more high-quality comments in the SENG2021 study compared to the SENG2020 study. This difference indicates that the customised nudges and visualisations for communication skills encouraged students to improve the quality of their comments. The number of videos students commented on in SENG2021 was three times higher than in the SENG2020 study. The SENG2021 participants made at least one comment on each video. The number of videos in which comments were rated in the SENG2021 study was significantly higher than in the SENG2020 study. Therefore, the nudges and visualisations helped students complete commenting and rating tasks for more videos in communication skills.

Table 6-4 Activities and interaction duration in the SENG2020/SENG2021 studies

	SENG2020 (N = 48)	SENG2021 (N = 49)	Significance
Sessions	5.47 (2.10)	6.67 (3.82)	t = -1.910, p = .06
Days	4.70 (1.70)	5.48 (2.59)	t = -1.756, p = .08
Unique videos	9.16 (2.16)	10.00 (.00)	t = -2.665, p = .01
Comments	9.39 (13.09)	29.55 (17.80)	t = -6.339, p < .001
Ratings	102.27 (153.25)	145.24 (280.87)	t = -.93, p = .17
Low-quality comments	5.53 (2.09)	5.24 (6.92)	t = .27, p = .78
High-quality comments	4.74 (1.69)	8.85 (6.40)	t = -4.33, p < .001
Commented videos	3.29 (3.45)	10.00 (0)	t = 13.44, p < .001
Rated videos	5.60 (4.45)	9.18 (2.76)	t = 4.74, p < .001

To investigate the effectiveness of Reminder nudges on using diverse aspects, aspects selected for comments on the tutorial and example videos in both studies were compared. Table 6-5 presents the distribution of aspects used in comments on tutorial and example videos in the SENG2020 and SENG2021 studies. The "*I didn't realise I wasn't doing this*" aspect was used more in the SENG2021 study than the SENG2020 study, and the use of the "*I like this point*" aspect decreased in the SENG2021 study. However, the chi-square test of homogeneity between studies and aspects used in comments on tutorial videos revealed no significant differences (Chi-square = 5.62, p = .13). On the other hand, the chi-square test showed a significant difference in the distribution of aspects used for comments on example videos (Chi-square = 21.52, p < .001) with a significant effect size of Phi = .144, p < .001. The post hoc analysis using z-test with Bonferroni correction showed that the percentage of *"Meeting contribution"* reduced significantly, while the proportion of the "*Receiving feedback*" aspect increased significantly in the SENG2021 study (p < .05). Therefore, nudges and visualisations helped students to use aspects more evenly in comments on example videos.

Table 6-5 Distribution of aspects used in comments on example videos

Aspects	SENG2020	SENG2021
Tutorial Videos		
I am rather good at this	24 (15.10%)	112 (16.20%)
I did/saw this in the past	35 (22.00%)	145 (21.00%)
I didn't realise I wasn't doing this	21 (13.20%)	143 (20.70%)
I like this point	79 (49.70%)	291 (42.10%)
Total	159	691
Example Videos		
Active listening	71 (24.3%)	199 (26.9%)
Giving feedback	28 (9.6%)	102 (13.8%)
Meeting contribution	104 (35.6%)	169 (22.9%)
Receiving feedback	13 (4.5%)	64 (8.7%)
Verbal communication	76 (26.0%)	205 (27.7%)
Total	294	739

To understand the differences between the comments in the SENG2020 and SENG2021 studies, the average of LIWC features and the domain-specific ratio was calculated for each student. The independent t-test showed no significant difference in the average domain-specific ratios between the two studies. Table 6-6 presents the average LIWC features which were statistically significantly different between the SENG2020 and SENG2021 studies. The average score for word per sentence, LIWC dictionary words, third-person pronouns ("she"/" he" and "they"), auxiliary verbs, negation (e.g., "never", "not") and differentiation words (e.g., "but", "otherwise", "instead") have increased significantly in the comments made on tutorial videos by students in the SENG2021 study compared to the SENG2020 study. These differences indicate that students wrote more dense comments, including more linguistic structures and differentiation when receiving nudges and visualisations. However, the SENG2020 participants had a significantly higher average score for using "I" pronouns and consequently made significantly more authentic comments. Nevertheless, the number of students who commented on tutorial videos in the SENG2021 study (49) was considerably higher than in the SENG2020 study (28).

For students who made comments on example videos in the SENG2021 study, the average score of authentic, LIWC dictionary words, "they" pronouns, impersonal pronouns, negation, common verbs (e.g., "talk", "listen"), comparisons (e.g., "best", "better") and differentiation were significantly higher compared to the SENG2020 study. This could indicate that students in the SENG2021 study showed more critical thinking and comparisons when watching example videos. However, the average analytics score for SENG2020 participants was significantly higher than the SENG2021 study. That could imply that students in the SENG2021 study made more here-and-now, narrative and disclosing comments on example videos. On the other hand, students in the SENG2020 study showed more formal and hierarchical thinking in comments on example videos (Pennebaker et al., 2015). However, the number of students who commented on example videos in the SENG2021 study (49) was considerably higher than in the SENG2020 study (21).

Table 6-6 Average LIWC features of comments made by students on tutorial/example videos

	SENG2020	SENG2021	Significance
Tutorial	(N = 28)	(N = 49)	
Authentic	53.78 (18.72)	45.26 (8.64)	t = 2.27, p < .05
Word per sentence	11.72 (4.45)	15.19 (4.56)	t = 3.23, p < .01
LIWC dictionary	88.63 (8.04)	92.26 (2.98)	t = 2.30, p < .05
I	4.92 (4.48)	2.86 (1.86)	t = 2.38, p < .05
She/he	.03 (.13)	.87 (.96)	t = 5.97, p < .001
They	1.18 (.95)	1.18 (.95)	t = 4.10, p < .001
Negate	1.52 (2.58)	3.16 (1.06)	t = 3.18, p < .01
Auxiliary verb	8.41 (5.15)	11.05 (2.33)	t = 2.56, p < .05
Common verbs	18.08 (5.96)	20.67 (3.22)	t = 2.12, p < .05
Differ	2.39 (2.73)	4.65 (1.27)	t = 4.11, p < .001
Example	(N = 21)	(N = 49)	
Authentic	26.32 (3.36)	46.44 (37.39)	t = 2.04, p < .05
Analytics	75.92 (31.76)	57.17 (31.63)	t = 2.26, p < .05
LIWC dictionary	87.14 (12.09)	93.15 (7.97)	t = 2.45, p < .05
They	.10 (.48)	2.30 (3.90)	t = 3.86, p< .001
Impersonal pronouns	2.96 (4.57)	6.74 (6.52)	t = 2.76, p< .01
Negate	.83 (2.22)	3.72 (5.18)	t = 3.27, p < .01
Common verbs	14.99 (12.92)	22.13 (13.43)	t = 2.06, p< .05
Compare	8.66 (10.07)	9.96 (7.84)	t = 2.67, p < .01
Differ	1.28 (3.67)	5.61 (5.99)	t = 3.69, p < .001

The previous study in the context of presentation skills showed that rating visualisations helped students use more diverse rating categories. Hence, I expected more diverse ratings in the SENG2021 study compared to the SENG2020 study. Table 6-7 presents the distributions of rating categories from the two studies. The chi-square test of homogeneity between study and rating options revealed significant differences in the distribution of used rating options (Chi-square = 564.56, p < .001) with a significant effect size of Phi = .27, p < .001. The post hoc analysis using the z-test with a Bonferroni correction showed that the usage of "I like this point" ratings decreased significantly in the SENG2021 study ($p < .05$). In contrast, the usage of "I didn't notice this", "I don't agree with this", and "I hadn't thought of this" ratings increased significantly ($p < .05$). Thus, rating visualisations encouraged students to use more diverse and reflective rating categories, similar to the previous study on presentation skills.

Table 6-7 Distribution of rating options used in the SENG2020 and SENG2021 studies

Ratings	SENG2020	SENG2021
I didn't notice this	160 (3.3%)	469 (6.6%)
I don't agree with this	102 (2.1%)	399 (5.6%)
I hadn't thought of this	234 (4.8%)	1041 (14.6%)
I like this point	3310 (67.4%)	3577 (50.3%)
This is useful for me	1104 (22.5%)	1631 (22.9%)
Total	4910	7117

Although there were more high-quality comments on tutorial videos in the SENG2021 study than in the SENG2020 study, the chi-square test showed no significant difference in the distribution of quality categories between these studies (Chi-square = 3.35, $p = .5$ and Phi = .06, $p = .5$). However, the median number of high-quality comments on tutorial videos in the SENG2021 study was 8, while this measurement for the SENG2020 study was 1. Conclusively, the comparison of comments quality in the SENG2020 and SENG2021 studies and the results presented in Table 6-4 to Table 6-7 provide evidence for supporting hypothesis H7.1.

The students in the SENG2021 and SENG2020 were categorised post hoc into three categories based on the ICAP framework. Students who watched videos but did not make any comments were classified as Passive. Active students were defined as those who wrote one (the median in the SENG2020 study) high-quality comment, and Constructive students as those who wrote two or more high-quality comments. Table 6-8 shows the distribution of these categories in the SENG2020 and SENG2021 studies. The chi-square test of homogeneity between years and ICAP categories revealed a significant difference (Chi-square = 29.46, $p < .001$) with an effect size (Phi) of .551 ($p < .001$). The post hoc analysis showed a significant increase in the percentage of Constructive students and a significant decrease in the percentage of Passive students in the SENG2021 study ($p < .001$). However, the number of Active students was not significantly different between the two studies. This result confirms hypothesis H7.2 and is consistent with the studies on nudges and visualisations in AVW-Space for presentation skills.

Table 6-8 Distribution of ICAP categories in the SENG2020 and SENG2021 studies

Study	Passive	Active	Constructive
SENG2020	20 (41.7%)	6 (12.5%)	22 (45.8%)
SENG2021	0	3 (6.1%)	46 (93.9%)

To address hypothesis H7.3, I compared the number of videos watched, comments and ratings and interactions with visualisations (hovering for longer than 5 seconds or clicking) between Constructive and Active groups in the SENG2021 study. The Mann-Whitney test showed a significant difference between Active and Constructive students only in the number of high- and low-quality comments on tutorial videos and interaction with rating visualisations. Considering the definition of these ICAP categories, it is expected to see Constructive students make significantly more high-quality comments than Active students. However, Constructive students also made significantly less low-quality comments than Active students. In addition, the Constructive group interacted with rating visualisations significantly more than the Active group. Nevertheless, the small size of the Active group prevents further investigation of the difference between the Active and Constructive students.

Table 6-9 Activities that Constructive students performed significantly differently from Active students in SENG2021

Activity	Active (N = 3)	Constructive (N = 46)	Significance
Low-quality tutorial comments	9.66 (5.03)	4.95 (6.97)	Mann-Whitney = 21.50, $p < .05$
High-quality tutorial comments	1 (0)	9.36 (6.27)	Mann-Whitney = .00, $p < .001$
Interaction with rating visualisation	2.66 (4.61)	16.5 (14.50)	Mann-Whitney = 23.00, $p < .05$

6.4.2 RQ7.2: Effects of Nudges and Visualisations on Learning

To investigate learning in the SENG2021 study, a pairwise t-test was conducted on pre-and post-study conceptual knowledge scores (Table 6-10). Although there was no significant correlation between CK1 and CK2 ($r = .27$, $p =.08$), the pairwise t-test revealed that CK2 was significantly higher than CK1 ($t = 2.823$, $p < .01$). Therefore, results presented in Table 6-10 confirm hypothesis H7.4. Unfortunately, the comparison of learning between the SENG2020 and SENG2021 studies was impossible since the SENG2020 study had no questionnaire for students' conceptual knowledge after the second phase (reviewing comments).

Table 6-10 Conceptual knowledge scores for SENG2021 participants who completed both surveys

	Mean	SD
CK1 (N = 40)	6.40	4.01
CK2 (N = 40)	8.62	4.26

Table 6-11 presents the conceptual knowledge of Active and Constructive students in the SENG2021 study. The Mann-Whitney test revealed no significant difference in CK1 (Mann-Whitney = 33.50, $p = .16$) between the Active and Constructive groups. The pairwise t-test on Constructive students' CK1 and CK2 showed a significant increase in their conceptual knowledge ($t = 2.70$, $p < .05$). Therefore, hypothesis H7.5 is confirmed. However, since there were only two Active students who completed Survey 2, it is not possible to compare CK2 scores.

Table 6-11 Conceptual knowledge of Active and Constructive students in the SENG2021 study

	CK1			CK2		
Category	N	Mean	SD	N	Mean	SD
Active	3	3.66	1.52	2	5.00	1.41
Constructive	46	7.02	4.73	38	8.81	4.28
Total	49	6.81	4.66	40	8.62	4.26

To address hypothesis H7.6, a model for the SENG2021 study (Figure 6-4) was developed. The goal of investigating the path model was mainly to compare it with previous studies on nudges and visualisations. The chi-square test (8.15) for this model (DF = 9, 21 estimated parameters) shows that the model's predictions were not statistically significantly different from the data ($p = .52$). The Comparative Fit Index (CFI) was 1, and the Root Mean Square Error of Approximation (RMSEA) was 0. Hence, the model is acceptable: CFI is greater than .9, and RMSEA is less than .06 (Hu & Bentler, 1999). The model indicates that more interaction with progress visualisation results in a higher number of times students watch videos, which causes a higher CK2 score ($p < .05$). In addition, the higher number of interactions with Personal Space visualisations results in receiving more Quality nudges ($p < .001$) and consequently making more high-quality comments ($p < .00$). Other links are all significant at $p < .05$ except High-quality → CK2 ($p = .34$).

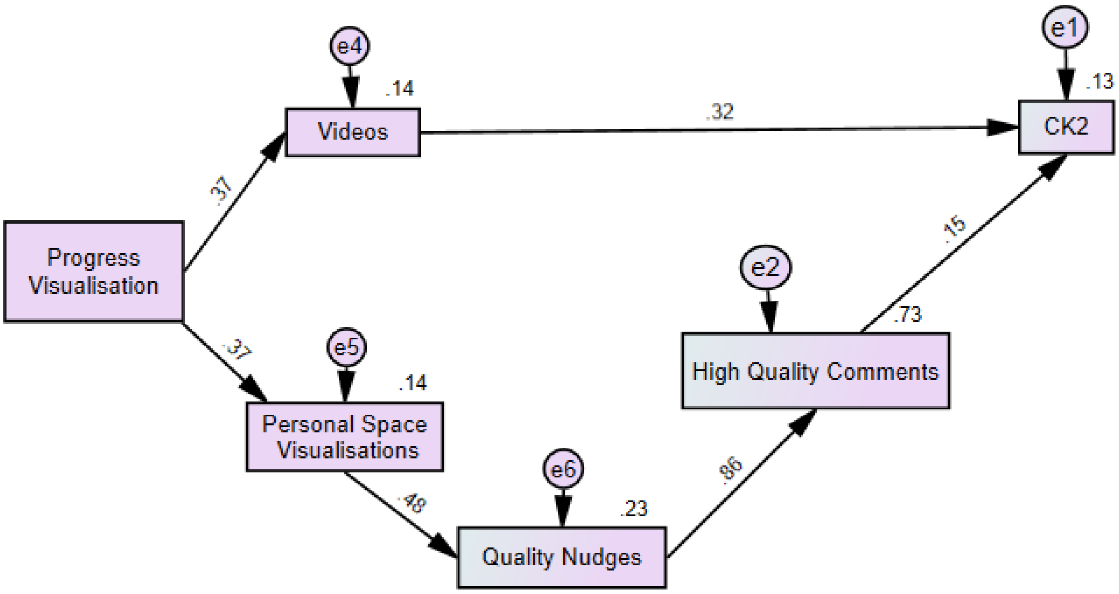

Figure 6-4 Path diagram for testing hypothesis H7.6: VLA and nudges increase the number of high-quality comments, watched videos and the post-study conceptual knowledge

To find the factors influencing learning for the SENG2021 participants, I developed a generalised linear model using the students' CK1, the number of watched videos, comments made and rated, nudges received, and interactions with visualisations in Personal Space to predict CK2 for students. The model performed well with Likelihood Ratio Chi-square = 31.46, $p < .001$ and AIC = 219.06. Table 6-12 present the parameter estimation of this generalised linear model. CK1 and the number of high-quality comments on tutorial videos were significant predictors of CK2, with the largest coefficients. The number of ratings made is also a significantly important factor, but it has a small coefficient. As seen with the path diagram, interactions with Personal Space visualisations and receiving nudges alone are not significant predictors of CK2. However, the interaction of nudges and visualisations with the number of high-quality comments on tutorial videos are important significant predictors of CK2. The negative coefficient for the interaction of high-quality comments and nudges on tutorial videos indicates that the effect of high-quality comments on CK2 will decrease as the number of received nudges gets larger. Although the generalised linear model and the path diagram provide evidence for supporting hypothesis H7.6, these results are not statistically reliable due to the small sample size.

Table 6-12 Parameter estimation of the generalised linear model for CK2

Parameter	β	Wald Chi-Square	Significance
(Intercept)	1.26	.20	.649
CK1	.90	15.07	.001
High-quality comments on tutorial videos	1.14	9.19	.002
Nudges on tutorial video	.03	.17	.679
Interaction with Personal Space visualisations	-.01	.29	.589
Ratings	-.01	7.45	.006
CK1 * High-quality comments on tutorial videos	-.11	26.36	.001
CK1 * Ratings	.002	11.72	.001
High-quality comments * Nudge on tutorial videos	-.022	4.23	.040
High-quality tutorial comments * Interaction with Personal Space visualisations	.005	4.26	.039
(Scale)	7.49		

6.4.3 RQ7.3: Engagement Differences in Studies on Presentation and Communication Skills

I was curious to know how students in two distinct domains interacted with AVW-Space. Thus, I used the logs of student interaction with Personal Space and Social Space from the SENG2021 study and ENG2021 studies (presented in Chapter 5). These studies provided the same type of support (visualisations and Reminder and Quality nudges), but the ENG2021 study was conducted with the first-year engineering students to train presentation skills. Since there were no Passive students in the SENG2021 study, only logs of Active and Constructive students in SENG2021 and ENG2021 studies were investigated. Then, ENA was applied to compare students' interactions in the SENG2021 and ENG2021 studies. Table 6-13 presents the codes used for generating the ENA networks.

Table 6-14 shows the frequencies of these codes in the logs for the SENG2021 and ENG2021 studies. As can be seen, ENGR2021 students had more interactions with the others' comments timeline visualisations, but this could be due to the higher number of comments provided in this visualisation. The ENG2021 students interacted with progress visualisation more than SENG2021 students. However, SENG2021 participants had significantly more frequent codes for ratings, since there were 1,464 comments selected for the rating phase in the SENG2021, but only 890 comments were chosen to be rated in the ENG2021 study. The mean ratings per student are 145.24 (SD = 280.87) in the SENG2021 study, while that measurement for the ENGR2021 study is 23.55 (SD = 52.26).

Table 6-13 Description of codes derived from event logs of SENG2021 and ENG2021 studies

Codes	Description	Example in the logs
Personal Space codes		
Video_load	Loading a video page	Pageload: Video Watch
Nudge	Receiving a nudge	Nudge
Comment	Making a comment	Comment created
Others_timeline_vis	Hovering over others comments timeline visualisation	Interaction with visualisation_name= Others_comments_timeline
Personal_timeline_vis	Hovering over personal comments timeline visualisation	Interaction with visualisation_name= Personal_comments_timeline
Progress_vis	Hovering over progress report visualisations	Interaction with visualisation_name= progress_report_visualisation
Nudge_vis	Clicking on nudges visualisations	Interaction with visualisation_name= nudge_visualisation
Social Space codes		
Review_load	Loading the rating page for a video	Video Review
Rating	Rating a comment	Rating_id=6, comment_id=1190
Rating_vis	Hovering over rating visualisations	Interaction with visualisation_name= rating_visualisation

Table 6-14 Frequency of events in SENG2021 and ENG2021 studies

Course	SENG2021 (N = 17,835)	ENG2021 (N = 48,562)
Comment	8.11 %	10.51 %
Others_timeline_vis	16.28 %	26.59 %
Personal_timeline_vis	1.77 %	1.93 %
Progress_vis	5.35 %	10 %
Nudge_vis	.79 %	1.75 %
Rating_vis	4.3 %	1.06 %
Nudge	12.18 %	17.64 %
Rating	44.24 %	17.93 %
Video_load	3.66 %	7.29 %
Review_load	3.28 %	5.25 %

The units of ENA were defined as all lines of data associated with a single study (ENG2021 or SENG2021), subsetted by ICAP Category and student_id. The conversations were defined as lines of data associated with a single student, subsetted by session. The stanza for the conversations was selected as 2 lines, meaning the codes of each line plus one previous line are aggregated within a given conversation to capture consequent events. In this model, networks were aggregated using a binary summation. The ENA model included the code presented in Table 6-13. I defined conversations as all lines of data associated with a single value of students_id, subsetted by session. The generated ENA model (Figure 6-5) had co-registration correlations of .97 (Pearson) and .96 (Spearman) for the first dimension and co-registration correlations of .97 (Pearson) and .97 (Spearman) for the second. These measures indicate strong goodness of fit between the visualisation (Figure 6-5) and the generated model. The two sample t-test assuming unequal variance showed the network for the ENG2021 study (mean = -.19, SD = .65, N = 326) was statistically significantly different from SENG2021 network (mean = 1.29, SD = 1.09, N = 49) along X-axis; t (53.23) = -9.35, p < .001, Cohen's d = 2.07). However, there was no significant difference along Y-axis between ENG2021 (mean = 0, SD = 1.12, N = 326) and SENG2021 networks (mean = 0, SD = 1.04, N = 49; (t (65.97) = 0, p = 1.00, Cohen's d = 0).

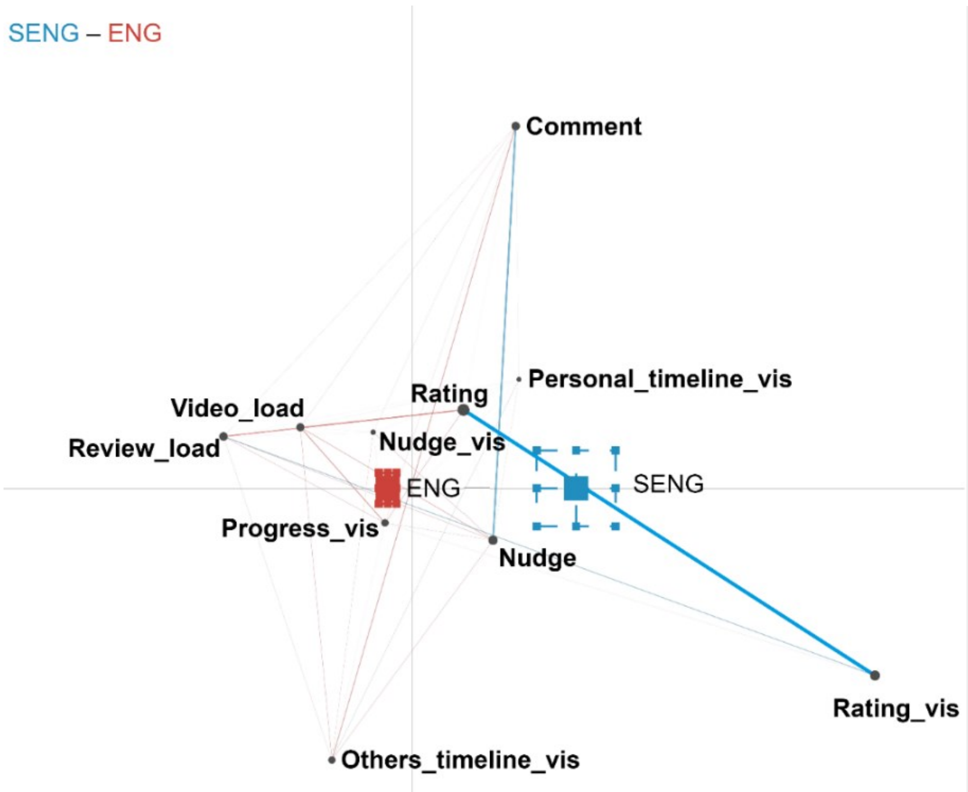

Figure 6-5 Difference network of SENG2021 and ENG2021 studies

As illustrated in Figure 6-5, the centroid for SENG2021 students is in the right half of the projected space where *Comments* and *Ratings*, *Rating_vis*, *Personal_timeline_vis* and *Nudge* are positioned. However, the centroid for the ENGR2021 network is on the left side of the projected space, where *Video_load*, *Review_load*, *Progress_vis*, *Nudge_vis* and *Others_timeline_vis* are located. As it can be seen, all co-occurrences were stronger for ENGR2021 participants except in *Rating_vis – Rating, Review_load – Rating_vis and Nudge – Comment*. In other words, *Rating_vis* and *Rating* co-occurred more often for SENG2021 students, which shows they were more interested in knowing peers' opinions on their comments than ENGR2021 students. Moreover, SENG2021 students were more likely to make a comment after receiving a nudge. At the same time, ENGR2021 participants benefited *Others_timeline_vis* and *Progress_vis* in order to comment. These differences in the use of nudges could be attributed to the familiarity of some SENG2021 students with AVW-Space through participating in the previous study on presentation skills. That is, some SENG2021 participants who used AVW-Space for communication skills had also participated in the previous study on Quality nudges for presentation skills in 2020 (presented in Chapter 4). However, I found no significant difference between SENG2021 students who experienced receiving nudges in the study on presentation skills (N = 20) and students who were new to the nudges (N = 29). Hence, the results of ENA confirms hypothesis H7.7.

6.5 Conclusions

In this chapter, the generalisability of nudges and visualisations was investigated for face-to-face meeting communication skills. First, the domain vocabulary for communication skills was extracted using data collected from the previous study on communication skills (the SENG2020 study), without nudges and visualisations. Then, the evaluation of quality assessment models on comments from the SENG2020 study showed that the quality assessment models are generalisable to this new domain and perform well. Next, AVW-Space was enhanced to allow teachers to customise the nudges and visualisation with regards to their selected videos and the target skill and domain. After customising nudges and visualisations for communication skills, the SENG2021 study was conducted to investigate the effectiveness of nudges and visualisations in learning communication skills. The results from this study showed that nudges and visualisations improved students' engagement in learning communication skills. Students who received nudges and visualisations watched more videos, made more high-quality comments, used more diverse aspects and rating options, commented on more videos, and rated the comments on more videos. Therefore, nudges and visualisations in communication skills boosted constructive behaviour and decreased the number of students who passively watched videos. In addition, students who made high-quality comments were more likely to have higher post-study conceptual knowledge. This chapter also presented the comparison of participants' interactions with nudges and visualisations in the presentation and communication skills domains. The ENA results revealed that students in the study on presentation skills benefited more from visualisations to comment. In contrast, students were more responsive to the nudges and made more ratings in the study on communication skills.

The main limitation of this study is the small sample size. The SENG202 course is a small course that runs in the second semester every year. In this study, I used the comments from a previous study on communication skills to generate domain vocabulary and customise examples for nudges and visualisations. In the case of cold start, where no data from a previous study is available for a new domain, vocabulary could be still extracted from YouTube auto-generated transcripts. However, in such a situation, the extracted vocabulary would be limited and require human evaluation and amendment, which could be time-consuming. In addition, curating an example for each nudge and video could be very demanding when the teacher makes many videos and nudges available for students. After conducting the SENG2021 study, I faced other challenges with the collected data and results. For instance, the small size of the Active group prevented further investigation on engagement and learning differences between ICAP categories. In addition, the lack of conceptual knowledge scores after the second phase of the SENG2020 study precluded the comparison of learning in students who received/did not receive nudges and visualisations. In addition, completing the last survey on the usefulness of AVW-Space and the cognitive load of commenting and rating was announced as voluntary due to COVID19 lockdown. This resulted in only a few responses to the survey. Moreover, I noticed that the increase in SENG2021 participants' engagement in terms of the number of watched videos and comments or ratings could also be also attributed to the increase in the course credit for participating in the study and the nudges visualisations. However, these factors were out of our control.

The future work for the generalisability of nudges and visualisations will firstly involve automating the extraction of domain vocabulary and enhancing the teacher interface to allow editing of the domain vocabulary. In addition, the examples for nudges can be automatically curated according to their predicted quality, provided the teacher has enough comments from previous studies or courses.

The study presented in this chapter investigated the generalisability of quality assessment models, nudges and visualisations only in the context of communication skills. However, communication skills have many overlapping concepts (e.g., verbal communication and body language) with presentation skills, for which the nudges were initially designed. Hence, further research is required to evaluate the quality assessment models, nudges and visualisations for other domains with fewer similarities to presentation skills. In addition, the investigation of nudges and visualisation for communication skills was only focused on the first and second phases of the studies (watching videos, commenting and rating). However, future research should investigate whether the nudges and visualisations can be applied to the third phase, in which the students watch and comment on their recorded team meetings and rate their teammates' comments.

7 Conclusions

This chapter first summarises this PhD research and its significant findings. Then, I discuss the limitations of this project and provide recommendations for future work.

7.1 Summary

This research aimed to enhance engagement in video-based learning by 1) developing personalised prompts for improving the quality of comments and 2) providing evocative visual analytics of the student model. Chapter 1 discussed my motivation for this research and the research questions followed by the proposed solution and research plan. The literature review (Chapter 2) underlined the most crucial challenge of maintaining and developing engagement in video-based learning, and presented several approaches to tackle these challenges and enhance engagement, such as visualisations and personalised prompts. However, research on the effectiveness of these approaches in video-based learning is still in the early stages.

I conducted my project in the context of AVW-Space, which is an online video-based learning platform that supports engagement via commenting and rating peers' comments. As discussed in Chapter 2, AVW-Space initially had simple prompts and visualisations. However, there was no support for improving the quality of comments and raising students' awareness of their learning progress. The review of previous studies on AVW-Space in Chapter 2 showed higher learning outcomes for students who write comments that show reflection and critical thinking. In addition, the students' feedback from previous studies highlighted the need for visualisations of the students' performance. Consequently, I proposed 1) Quality nudges that encourage students to write high-quality comments showing critical thinking and self-reflection, and 2) visualisations of the student model to facilitate monitoring and regulating the learning process for students. I applied quantitative and ethnographic analysis methods to examine the impact of the Quality nudges and visualisations on students' learning and engagement.

In chapter 3, I proposed and evaluated two quality schemes for assessing comments in AVW-Space. Then, I automated the quality assessment of comments by developing cost-sensitive classifiers. Following that, I evaluated the performance and generalisability of these models on unseen data. Next, the quality assessment models, and student profiles were used for Quality nudges, which encourage students to improve the quality of their comments by triggering critical thinking, self-reflection and self-regulation.

In Chapter 4, I investigated the effectiveness of Quality nudges by conducting a study with the new version of AVW-Space. The study revealed that Quality nudges effectively enhance engagement; there was a significant increase in the quantity and quality of comments, the time spent on AVW-Space, and the number of ratings made. ENA on the logs of students' interactions showed that low-performing students ignored nudges and continued watching videos rather than responding to nudges by commenting. This study also showed a significant increase in the conceptual knowledge of participants who made high-quality comments and received more Quality nudges. In addition, students who received Quality nudges perceived less cognitive load during commenting and rating. However, the analysis of native English speakers and students who speak English as a Foreign Language (EFL) showed that nudges were not equally helpful for these two groups. Most EFL students only watched the videos without writing comments, even after receiving nudges. Furthermore, EFL students had lower conceptual knowledge scores before and after the study compared to Native students. These differences could be attributed to language challenges and differences in the learning strategies of these groups.

In Chapter 5, I proposed the integration of various visual learning analytics into AVW-Space to boost engagement and learning. These visualisations were designed to assist students in monitoring and controlling their learning progress. I investigated the effectiveness of these visualisations by conducting a study with the enhanced version of AVW-Space. The study revealed that the visualisations effectively enhance engagement in VBL. The study with the visualisations showed an increase in the number of watched videos, comments, diversity of used rating options and time spent on AVW-Space. Furthermore, students who received the visual learning analytics in addition to nudges showed more constructive engagement and made high-quality comments compared to the ones who received only the nudges. The EFL students who received the visualisations and nudges showed more constructive engagement. The participants who made more high-quality comments also interacted more with the visualisations and significantly increased their conceptual knowledge. The study with the visualisations also revealed less cognitive load perceived in rating, and students found visualisations helpful.

Since the Quality nudges and visualisations were evaluated only in the context of presentation skills, I investigated their generalisability for face-to-face meeting communication skills in Chapter 6. After extracting the domain vocabulary and evaluating the generalisability of the quality assessment models for communication skills, nudges and visualisations were customised for this domain. Next, a study was conducted to investigate the effectiveness of nudges and visualisations in learning communication skills. This study showed that nudges and visualisations enhanced students' engagement in learning communication skills. Students who received nudges and visualisations watched more videos, made more high-quality comments, used more diverse aspects and rating options. Hence, nudges and visualisations in communication skills boosted constructive behaviour. In addition, students who wrote high-quality comments were more likely to have higher conceptual knowledge after the intervention. Chapter 6 also compared interactions with nudges and visualisations in the presentation and communication skills domains. The ENA results indicated that students in the study on presentation skills benefitted more from visualisations. However, students were more responsive to the nudges and made more ratings in the study on communication skills.

7.2 Limitations

One of the limitations of this research is that the population in all studies was exclusively from an engineering background. Thus, the enhanced version of AVW-Space with nudges and visualisations must be studied on students from a non-engineering background. Another shortcoming is the small sample size of the study conducted on communication skills. However, the main challenge of this research is quantifying the learning of transferable skills. In all studies conducted on AVW-Space, learning was measured by counting the number of domain-specific terms students listed in the surveys. This method was chosen due to its simplicity in measuring learning in large classes. However, this type of assessment is memory-based and does not represent the learner’s transferable skills. Therefore, a more sophisticated approach is required to assess the students’ presentation skills before and after using AVW-Space. Another drawback in all three studies is the ENA Webtool, which did not show the direction of co-occurrence to clarify their sequence. Moreover, some interactions such as skipping the video and hovering over nudges, were not logged in the system for ENA to analyse.

In the existing implementation, all nudges are evaluated upon any model update. This could result in unacceptably slow performance of the nudge engine when the number of nudges increases. In addition, choosing the priority for each nudge could be a complicated task as the number of nudges grows. Another drawback of the current implementation is that adding and editing nudges must be implemented as a class which require programming. Therefore, an authoring tool is needed to facilitate adding and editing nudges. Moreover, the examples for nudges are static. In other words, if a student receives a specific nudge twice for a video, the message of the nudge would be the same and does not offer new information to the student. Therefore, nudges should be more context-aware and relevant to the part of the video the student is watching.

In the enhanced version of AVW-Space, all students were provided with the same type of visualisations. However, the study on visualisations showed students have different preferences about the visualisations they would like to see and perceive the usability of each visualisation differently. Hence, the visibility of visualisations should be tailored to the student's preferences and learning strategies. One challenge is that showing others' comments timeline requires comments from previous studies, which are not available in the case of cold start. Another drawback of the comment timeline is that the teachers must manually select the comments shown, which could be tedious for a large class. The major challenge in studying the effectiveness of visualisations is determining how to accurately assess the hovering with visualisations and identify which ones were intentional and focused. In this research, I analysed hovering interactions continued for longer than 5 seconds and discarded interactions with shorter duration as I considered them unintentional. However, the duration of interaction is not a comprehensive criterion for determining intentional hoverings, and a more sophisticated approach is required for analysing these interactions.

Although nudges and visualisations were shown to be generalisable for another soft skill, extracting the domain vocabulary and customising nudges and visualisations requires comments from previous studies. In addition, curating an example for each nudge and video could be demanding for the teachers. A challenge in the study on the generalisability of quality assessment models, nudges and visualisations in communication skills is that this domain has overlapping concepts with presentation skills (e.g., body language and verbal communication). Hence, the generalisability of nudges and visualisations presented in this research could be partially attributed to the similarity of this domain to presentation skills, for which the nudges and quality assessment models were designed initially. Another limitation of this study is insufficient feedback on the usefulness of AVW-Space and cognitive load of commenting and rating since the COVID19 lockdown affected the participation of students in this study. The following section suggests future work for addressing the issues discussed.

7.3 Future Work

As discussed earlier, the nudges and visualisations must be evaluated on non-engineering students and in the context of other skills having less overlap with presentation skills. In addition, a better instrument for evaluating students' knowledge should be devised to assess learning outcomes more thoroughly. Although nudges and visualisations trigger students' self-reflection and self-regulation to promote engagement, the impact of these supports on improving self-reflection and self-regulation skills was not investigated in this research. Future research could assess the self-reflection and self-regulation skills before and after the interventions.

The performance of the quality assessment models could be enhanced further by a more profound feature engineering and retraining the models on the larger dataset gathered from the two studies on Quality nudges and visualisations. In addition, future research could explore beyond LIWC and domain-specific features by examining word-embedding approaches and using larger corpus from other platforms (e.g., YouTube comments). An authoring tool could be developed in future to allow teachers to customise or add new nudges by defining them in a simple formal language. The Quality nudges could also be enhanced by considering those students who did not respond to nudges. For example, if a student constantly makes comments repeating the video content, giving the same general nudge for critical thinking might be ineffective. Instead, the learner might need more explicit and detailed instruction for thinking critically. As some negative feedback on nudges noted, nudges messages are generic and static. Hence, the future enhancement of the nudges could contain more reactive messages with more specific instructions aligned with the current topic of the video and followed by a variety of example comments. To provide various example comments for nudges, the quality assessment models could be leveraged for selecting high-quality comments. The nudges could also be adapted to EFL students in future to offer simplified instructions with a longer displaying period.

As students suggested, more explanation and interactivity (e.g., filtering, searching) should be added for the visualisations in future to increase their usability. In addition, the future version of AVW-Space could offer a visibility option for visualisations that triggers social comparison. Future work could also involve further investigation of students' learning strategies and tailoring visualisations to the students' behaviours. New nudges could also be added in the future version to encourage the learners to use specific visualisations according to their behaviour and needs. In addition, future research could explore solutions for automating the selection of comments to display in the comment timeline visualisation. Future studies should also explore eye-gaze analysis and think-aloud strategies to better understand how the visualisations drive students to regulate their learning.

The future work for the generalisability of nudges and visualisations could involve automating the extraction of domain vocabulary and enhancing the teacher interface to allow editing of the domain vocabulary. Future research could also investigate whether and how the nudges and visualisations can be applied to the third phase, in which the students watch and comment on their recorded team meetings and rate their teammates' comments. Nudges for meeting comments could be a combination of nudges for examples and tutorial comments or a new set of nudges designed for this particular phase.

Although the focus of this research was mainly on the Personal Space of AVW-Space, new nudges could also be integrated into the reviewing task in future research. For example, nudges could encourage using diverse rating options or give positive feedback when students rate a good-quality comment. To provide such support, the record of ratings students made should be maintained. The efficacy of Social Space could be further enhanced by allowing more interactivity in the comments list to rate. For example, future versions could allow students to group and filter comments by their quality or aspect. Since students usually rate the top comments, students whose comments are in the middle or bottom of the list might not receive many ratings, so the rating visualisations might not be useful for them. Thus, an approach for reshuffling the ratings intelligently could be developed in future to show unrated good quality comments and allow all comments to have the chance to be rated.

7.4 Contributions

The main contribution of this research is addressing the need for personalisation and feedback in VBL, since literature warns that the lack of these factors causes low engagement. This research contributes to the development of intelligent learning environments which provide personalised interventions in the form of nudges to foster good comment writing behaviours during video-based learning. The findings of this research emphasise the effectiveness of personalised SRL scaffolds for achieving high learning outcomes. As the literature suggests, meta-cognitive activities are essential for constructive learning. However, adapting meta-cognitive activities to the student’s learning strategy and progress enhances learning outcomes as literature shows different formats of support suited to different students. The interventions proposed in this research can be applied to other domains where critical thinking and self-reflection are required. Another contribution of this research is eliciting the requirements for improving inclusiveness in computer-assisted learning environments and improving equity in the learning experience and outcomes for non-native English speakers. The important findings from this research encourage researchers to investigate the equity for non-native English speakers in other platforms. This research also contributes to combining ENA and quantitative analysis to better understand student behaviour in computer-assisted educational systems to discover new ways of supporting students' needs.

Another contribution of this research is using student-facing visual learning analytics in video-based learning platforms to boost engagement and learning. My findings confirm the literature on the effectiveness of visual learning analytics in encouraging students to monitor and control their learning. The proposed visualisations in this thesis are a form of SRL support since they offer feedback on the learner's progress and allow the learner to reflect on their learning and take action according to their progress. The nudges and visualisations proposed in this research could be applicable in any other video-based learning platform that supports commenting.

References

Agarwala, M., Hsiao, Ih., Chae, H. S., & Natriello, G. (2012). Vialogues: Videos and Dialogues Based Social Learning Environment. *2012 IEEE 12th International Conference on Advanced Learning Technologies*, 629–633. https://doi.org/10.1109/ICALT.2012.127

Aguilar, S., Karabenick, S. A., Teasley, S. D., & Baek, C. (2021). Associations between Learning Analytics Dashboard Exposure and Motivation and Self-regulated Learning. *Computers & Education*, *162*, 104085. https://doi.org/10.1016/j.compedu.2020.104085

Aguilar, S., Lonn, S., & Teasley, S. D. (2014). Perceptions and Use of an Early Warning System during a Higher Education Transition Program. *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, 113–117. https://doi.org/10.1145/2567574.2567625

Ahmet, A., Gamze, K., Rustem, M., & Sezen, K. A. (2018). Is Video-Based Education an Effective Method in Surgical Education? A Systematic Review. *Journal of Surgical Education*, *75*(5), 1150–1158. https://doi.org/10.1016/j.jsurg.2018.01.014

Alavi, M. (1994). Computer-Mediated Collaborative Learning: An Empirical Evaluation. *MIS Quarterly*, *18*(2), 159–174. https://doi.org/10.2307/249763

Almatrafi, O., & Johri, A. (2019). Systematic Review of Discussion Forums in Massive Open Online Courses (MOOCs). *IEEE Transactions on Learning Technologies*, *12*(3), 413–428. https://doi.org/10.1109/TLT.2018.2859304

Alten, D. C. D. van, Phielix, C., Janssen, J., & Kester, L. (2020). Self-regulated learning support in flipped learning videos enhances learning outcomes. *Computers & Education*, *158*, 104000. https://doi.org/10.1016/j.compedu.2020.104000

Amarasinghe, I., Hernández-Leo, D., & Jonsson, A. (2019). Data-informed design parameters for adaptive collaborative scripting in across-spaces learning situations. *User Modeling and User-Adapted Interaction*, *29*(4), 869–892. https://doi.org/10.1007/s11257-019-09233-8

Anthony, S., & Garner, B. (2016). Teaching Soft Skills to Business Students: An Analysis of Multiple Pedagogical Methods. *Business and Professional Communication Quarterly*, *79*(3), 360–370. https://doi.org/10.1177/2329490616642247

Apoki, U. C., & Crisan, G. C. (2019). Employing Software Agents and Constructivism to make Massive Open Online Courses more student-oriented. *2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 1–8. https://doi.org/10.1109/ECAI46879.2019.9042001

Arbogast, C. A., & Montfort, D. (2016). Applying natural language processing techniques to an assessment of student conceptual understanding. *ASEE Annual Conference and Exposition, Conference Proceedings*, *2016*, 13. https://doi.org/10.18260/p.26262

Azevedo, R., Landis, R. S., Feyzi-Behnagh, R., Duffy, M., Trevors, G., Harley, J. M., Bouchet, F., Burlison, J., Taub, M., Pacampara, N., Yeasin, M., Rahman, A. K. M. M., Tanveer, M. I., & Hossain, G. (2012). The Effectiveness of Pedagogical Agents' Prompting and Feedback in Facilitating Co-adapted Learning with MetaTutor. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 212–221). Springer. https://doi.org/10.1007/978-3-642-30950-2_27

Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning*, *3*(1), 39–58. https://doi.org/10.1007/s11409-007-9009-6

Boateng, R., Boateng, S. L., Awuah, R. B., Ansong, E., & Anderson, A. B. (2016). Videos in Learning in Higher Education: Assessing Perceptions and Attitudes of Students at the University of Ghana. *Smart Learning Environments*, *3*(1), 8. https://doi.org/10.1186/s40561-016-0031-5

Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open Learner Models and Learning Analytics Dashboards: A Systematic Review. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 41–50. https://doi.org/10.1145/3170358.3170409

Bouchet, F., Harley, J. M., & Azevedo, R. (2016). Can Adaptive Pedagogical Agents' Prompting Strategies Improve Students' Learning and Self-Regulation? In A. Micarelli, J. Stamper, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 368–374). Springer International Publishing.

Broos, T., Peeters, L., Verbert, K., Van Soom, C., Langie, G., & De Laet, T. (2017). Dashboard for Actionable Feedback on Learning Skills: Scalability and Usefulness. In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies. Technology in Education* (pp. 229–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-58515-4_18

Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R., Zadorozhny, V., & Durlach, P. J. (2016). Open Social Student Modeling for Personalized Learning. *IEEE Transactions on Emerging Topics in Computing*, *4*(3), 450–461. https://doi.org/10.1109/TETC.2015.2501243

Bull, S., & Kay, J. (2010). Open Learner Models. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems* (pp. 301–322). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14363-2_15

Chadha, D. (2006). A curriculum model for transferable skills development. *Engineering Education*, *1*(1), 19–24. https://doi.org/10.11120/ened.2006.01010019

Charleer, S., Moere, A. V., Klerkx, J., Verbert, K., & De Laet, T. (2018). Learning Analytics Dashboards to Support Adviser-Student Dialogue. *IEEE Transactions on Learning Technologies*, *11*(3), 389–399. https://doi.org/10.1109/TLT.2017.2720670

Chatti, M. A., Marinov, M., Sabov, O., Laksono, R., Sofyan, Z., Fahmy Yousef, A. M., & Schroeder, U. (2016). Video Annotation and Analytics in CourseMapper. *Smart Learning Environments*, *3*(1), 10. https://doi.org/10.1186/s40561-016-0035-1

Chen, Y., Gao, Q., Yuan, Q., & Tang, Y. (2019). Facilitating Students' Interaction in MOOCs through Timeline-Anchored Discussion. *International Journal of Human–Computer Interaction*, *35*(19), 1781–1799. https://doi.org/10.1080/10447318.2019.1574056

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Chin-Yuan Lai, Li-Ju Chen, Yung-Chin Yen, & Kai-Yin Lin. (2020). Impact of video annotation on undergraduate nursing students' communication performance and commenting behaviour during an online peer-assessment activity. *Australasian Journal of Educational Technology*, *36*(2), 71–88. https://doi.org/10.14742/ajet.4341

Chiu, P.-S., Chen, H.-C., Huang, Y.-M., Liu, C.-J., Liu, M.-C., & Shen, M.-H. (2018). A video annotation learning approach to improve the effects of video learning. *Innovations in Education and Teaching International*, *55*(4), 459–469. https://doi.org/10.1080/14703297.2016.1213653

Chou, C.-Y., Tseng, S.-F., Chih, W.-C., Chen, Z.-H., Chao, P.-Y., Lai, K. R., Chan, C.-L., Yu, L.-C., & Lin, Y.-L. (2017). Open Student Models of Core Competencies at the Curriculum Level: Using Learning Analytics for Student Reflection. *IEEE*

Transactions on Emerging Topics in Computing, *5*(1), 32–44. https://doi.org/10.1109/TETC.2015.2501805

Cohen, W. W. (1995). Fast Effective Rule Induction. In A. Prieditis & S. Russell (Eds.), *Machine Learning Proceedings 1995* (pp. 115–123). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-377-6.50023-2

Colasante, M., & Colasante, M. (2011). Using Video Annotation to Reflect on and Evaluate Physical Education Pre-service Teaching Practice. *Australasian Journal of Educational Technology*, *27*(1), 66–88.

Conkey, C. A., Bowers, C., Cannon-Bowers, J., & Sanchez, A. (2013). Machinima and Video-Based Soft-Skills Training for Frontline Healthcare Workers. *Games for Health Journal*, *2*(1), 39–43. https://doi.org/10.1089/g4h.2012.0063

Corrin, L., & de Barba, P. (2014). Exploring students' interpretation of feedback delivered through learning analytics dashboards. *Rhetoric and Reality: Critical Perspectives on Educational Technology. Proceedings Ascilite Dunedin*, 629–633.

Cronin, M. W., & Cronin, K. A. (1992). Recent empirical studies of the pedagogical effects of interactive video instruction in "soft skill" areas. *Journal of Computing in Higher Education*, *3*(2), 53. https://doi.org/10.1007/BF02942356

Crossley, S., Kim, M., Allen, L., & McNamara, D. (2019). Automated Summarization Evaluation (ASE) Using Natural Language Processing Tools. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education* (pp. 84–95). Springer International Publishing.

Crossley, S., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. (2015). Language to Completion: Success in an Educational Data Mining Massive Open Online Class. *Proceedings of the 8th International Conference on Educational Data Mining*, 388–391. https://eric.ed.gov/?id=ED560771

Cummins, S., Beresford, A. R., & Rice, A. (2016). Investigating Engagement with In-Video Quiz Questions in a Programming Course. *IEEE Transactions on Learning Technologies*, *9*(1), 57–66. https://doi.org/10.1109/TLT.2015.2444374

Cunningham-Nelson, S., Mukherjee, M., Goncher, A., & Boles, W. (2018). Text analysis in education: A review of selected software packages with an application for analysing students' conceptual understanding. *Australasian Journal of Engineering Education*, *23*(1), 25–39. https://doi.org/10.1080/22054952.2018.1502914

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. JSTOR. https://doi.org/10.2307/249008

Dawson, S., Macfadyen, L., Risko, E. F., Foulsham, T., & Kingstone, A. (2012). Using technology to encourage self-directed learning: The Collaborative Lecture Annotation System (CLAS). *Annual Conference of the Australian Society for Computers in Learning in Tertiary Education (ASCILITE 2012)*, 25–28.

Demetriadis, S., Karakostas, A., Tsiatsos, T., Caballé, S., Dimitriadis, Y., Weinberger, A., Papadopoulos, P. M., Palaigeorgiou, G., Tsimpanis, C., & Hodges, M. (2018). Towards Integrating Conversational Agents and Learning Analytics in MOOCs. In L. Barolli, F. Xhafa, N. Javaid, E. Spaho, & V. Kolici (Eds.), *Advances in Internet, Data & Web Technologies* (pp. 1061–1072). Springer International Publishing. https://doi.org/10.1007/978-3-319-75928-9_98

Dimitrova, V., & Brna, P. (2016). From Interactive Open Learner Modelling to Intelligent Mentoring: STyLE-OLM and Beyond. *International Journal of Artificial Intelligence in Education*, *26*(1), 332–349. https://doi.org/10.1007/s40593-015-0087-3

Dimitrova, V., & Mitrovic, A. (2021). Choice Architecture for Nudges to Support Constructive Learning in Active Video Watching. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00263-1

Dimitrova, V., Mitrovic, A., Piotrkowicz, A., Lau, L., & Weerasinghe, A. (2017). Using Learning Analytics to Devise Interactive Personalised Nudges for Active Video Watching. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 22–31.

Dimitrova, V., Steiner, C., Despotakis, D., Brna, P., Ascolese, A., Pannese, L., & Albert, D. (2013). *Crowdsourcing for Evaluating a Simulated Learning Environment for Interpersonal Communication and Cultural Awareness*. Workshop on Culturally-aware Technology Enhanced Learning (CulTEL 2013).

Dodson, S., Roll, I., Fong, M., Yoon, D., Harandi, N. M., & Fels, S. (2018). Active Viewing: A Study of Video Highlighting in the Classroom. *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, 237–240. https://doi.org/10.1145/3176349.3176889

Ebied, M., Kahouf, S. A., & Rahman, S. A. A. (2016). Effectiveness of Using Youtube in Enhance the Learning of Computer in Education Skills in Najran University. *Nternational Interdisciplinary Journal of Education*, *5*(3), 620–625.

Evi-Colombo, A., Cattaneo, A., & Bétrancourt, M. (2020). Technical and Pedagogical Affordances of Video Annotation: A Literature Review. *Journal of Educational Multimedia and Hypermedia*, *29*(3), 193–226.

Ez-zaouia, M., Tabard, A., & Lavoué, E. (2020). Emodash: A dashboard supporting retrospective awareness of emotions in online learning. *International Journal of Human-Computer Studies*, *139*, 102411. https://doi.org/10.1016/j.ijhcs.2020.102411

Frank, E., & Hall, M. (2001). A Simple Approach to Ordinal Classification. In L. De Raedt & P. Flach (Eds.), *Machine Learning: ECML 2001* (pp. 145–156). Springer Berlin Heidelberg.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Galster, M., Mitrovic, A., & Gordon, M. (2018). Toward enhancing the training of software engineering students and professionals using active video watching. *Proceedings of the 40th International Conference on Software Engineering: Software Engineering Education and Training*, 5–8.

Gamage, D., Perera, I., & Fernando, S. (2020). Exploring MOOC User Behaviors Beyond Platforms. *International Journal of Emerging Technologies in Learning (IJET)*, *15*(8), 161–179.

Gasevic, D., Jovanovic, J., Pardo, A., & Dawson, S. (2017). Detecting Learning Strategies with Analytics: Links with Self-reported Measures and Academic Performance. *Journal of Learning Analytics*, *4*(2), 113-128-113–128. https://doi.org/10.18608/jla.2017.42.10

Gašević, D., Mirriahi, N., & Dawson, S. (2014). Analytics of the Effects of Video Use and Instruction to Support Reflective Learning. *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, 123–132. https://doi.org/10.1145/2567574.2567590

Giannakos, M. N., Sampson, D. G., & Kidziński, Ł. (2016). Introduction to smart learning analytics: Foundations and developments in video-based learning. *Smart Learning Environments*, *3*(1), 1–9. https://doi.org/10.1186/s40561-016-0034-2

Gilboy, M. B., Heinerichs, S., & Pazzaglia, G. (2015). Enhancing Student Engagement Using the Flipped Classroom. *Journal of Nutrition Education and Behavior*, *47*(1), 109–114. https://doi.org/10.1016/j.jneb.2014.08.008

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *The Elementary School Journal*, *115*(2), 210–229. https://doi.org/10.1086/678293

Grann, J., & Bushway, D. (2014). Competency map: Visualizing student learning to promote student success. *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, 168–172. https://doi.org/10.1145/2567574.2567622

Guerra, J., Hosseini, R., Somyurek, S., & Brusilovsky, P. (2016). An Intelligent Interface for Learning Content: Combining an Open Learner Model and Social Comparison to Support Self-Regulated Learning and Engagement. *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 152–163. https://doi.org/10.1145/2856767.2856784

Haagsman, M. E., Scager, K., Boonstra, J., & Koster, M. C. (2020). Pop-up Questions Within Educational Videos: Effects on Students' Learning. *Journal of Science Education and Technology*, *29*(6), 713–724. https://doi.org/10.1007/s10956-020-09847-3

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hecking, T., Dimitrova, V., Mitrovic, A., & Hoppe, H. U. (2017). Using Network-Text Analysis to Characterise Learner Engagement in Active Video Watching. *ICCE 2017 Main Conference Proceedings*, 326–335.

Heilman, M., & Madnani, N. (2015). The Impact of Training Data on Automated Short Answer Scoring Performance. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 81–85. https://doi.org/10.3115/v1/W15-0610

Hetzner, S., Steiner, C. M., Dimitrova, V., Brna, P., & Conlan, O. (2011). Adult Self-regulated Learning through Linking Experience in Simulated and Real World: A Holistic Approach. In C. D. Kloos, D. Gillet, R. M. Crespo García, F. Wild, & M. Wolpers (Eds.), *Towards Ubiquitous Learning* (pp. 166–180). Springer Berlin Heidelberg.

Hooshyar, D., Pedaste, M., Saks, K., Leijen, Ä., Bardone, E., & Wang, M. (2020). Open learner models in supporting self-regulated learning in higher education: A systematic literature review. *Computers & Education*, *154*, 103878. https://doi.org/10.1016/j.compedu.2020.103878

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hulsman, R. L., Harmsen, A. B., & Fabriek, M. (2009). Reflective teaching of medical communication skills with DiViDU: Assessing the level of student reflection on recorded consultations with simulated patients. *Patient Education and Counseling*, *74*(2), 142–149. https://doi.org/10.1016/j.pec.2008.10.009

Hulsman, R. L., & van der Vloodt, J. (2015). Self-evaluation and peer-feedback of medical students' communication skills using a web-based video annotation system. Exploring content and specificity. *Patient Education and Counseling*, *98*(3), 356–363. https://doi.org/10.1016/j.pec.2014.11.007

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced Data—Recommendations for the Use of Performance Metrics. *2013 Humaine Association Conference on*

Affective Computing and Intelligent Interaction, 245–251. https://doi.org/10.1109/ACII.2013.47

Joksimović, S., Dowell, N., Gašević, D., Mirriahi, N., Dawson, S., & Graesser, A. C. (2019). Linguistic characteristics of reflective states in video annotations under different instructional conditions. *Computers in Human Behavior*, *96*, 211–222. https://doi.org/10.1016/j.chb.2018.03.003

Jung, Y., & Wise, A. F. (2020). How and How Well Do Students Reflect? Multi-Dimensional Automated Reflection Assessment in Health Professions Education. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 595–604. https://doi.org/10.1145/3375462.3375528

Karumbaiah, S., Baker, R. S., Barany, A., & Shute, V. (2019). Using Epistemic Networks with Automated Codes to Understand Why Players Quit Levels in a Learning Game. In B. Eagan, M. Misfeldt, & A. Siebert-Evenstone (Eds.), *Advances in Quantitative Ethnography* (pp. 106–116). Springer International Publishing. https://doi.org/10.1007/978-3-030-33232-7_9

Kim, J., Guo, P. J., Cai, C. J., Li, S.-W. (Daniel), Gajos, K. Z., & Miller, R. C. (2014). Data-Driven Interaction Techniques for Improving Navigation of Educational Videos. *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, 563–572. https://doi.org/10.1145/2642918.2647389

Kleftodimos, A., & Evangelidis, G. (2016). Using open source technologies and open internet resources for building an interactive video based learning environment that supports learning analytics. *Smart Learning Environments*, *3*(1), 9. https://doi.org/10.1186/s40561-016-0032-4

Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., & Pineau, J. (2020). Automated Personalized Feedback Improves Learning Gains in An Intelligent Tutoring System. In

I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 140–146). Springer International Publishing.

Kopolovich, O. (2020). Learning soft skills in the digital age: Challenges and insights from development and teaching “Negotiation Management” MOOC. *Online Journal of Applied Knowledge Management (OJAKM)*, *8*(2), 91–106. https://doi.org/10.36965/OJAKM.2020.8(2)91-106

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Krippendorff’s Alpha. (2010). In N. Salkind, *Encyclopedia of Research Design* (p. 670). SAGE Publications, Inc. https://doi.org/10.4135/9781412961288.n206

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. JSTOR. https://doi.org/10.2307/2529310

Lim, L., Dawson, S., Joksimovic, S., & Gašević, D. (2019). Exploring Students’ Sensemaking of Learning Analytics Dashboards: Does Frame of Reference Make a Difference? *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 250–259. https://doi.org/10.1145/3303772.3303804

Liu, C., Kim, J., & Wang, H.-C. (2018). ConceptScape: Collaborative Concept Mapping for Video Learning. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Liu, C., Yang, C.-L., Williams, J. J., & Wang, H.-C. (2019). NoteStruct: Scaffolding Note-taking while Learning from Online Videos. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312878

Liu, M., Shum, S. B., Mantzourani, E., & Lucas, C. (2019). Evaluating Machine Learning Approaches to Classify Pharmacy Students' Reflective Statements. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education* (pp. 220–230). Springer International Publishing.

Lonn, S., Aguilar, S., & Teasley, S. D. (2015). Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. *Computers in Human Behavior*, *47*, 90–97. https://doi.org/10.1016/j.chb.2014.07.013

Marquart, C. L., Hinojosa, C., Swiecki, Z., Eagan, B., & Shaffer, D. (2018). *Epistemic Network Analysis (Version 1.7.0) [Software]*. http://app.epistemicnetwork.org

Martín-Monje, E., Castrillo, M. D., & Mañana-Rodríguez, J. (2018). Understanding online interaction in language MOOCs through learning analytics. *Computer Assisted Language Learning*, *31*(3), 251–272. https://doi.org/10.1080/09588221.2017.1378237

Matcha, W., Uzir, N. A., Gašević, D., & Pardo, A. (2020). A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. *IEEE Transactions on Learning Technologies*, *13*(2), 226–245. https://doi.org/10.1109/TLT.2019.2916802

Mayer, R. E. (2021). Evidence-Based Principles for How to Design Effective Instructional Videos. *Journal of Applied Research in Memory and Cognition*, *10*(2), 229–240. https://doi.org/10.1016/j.jarmac.2021.03.007

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499–515. https://doi.org/10.3758/s13428-012-0258-1

Mejia, C., Florian, B., Vatrapu, R., Bull, S., Gomez, S., & Fabregat, R. (2017). A Novel Web-Based Approach for Visualization and Inspection of Reading Difficulties on University

Students. *IEEE Transactions on Learning Technologies*, *10*(1), 53–67. https://doi.org/10.1109/TLT.2016.2626292

Mendoza, G. L. L., Caranto, L. C., & David, J. J. T. (2015). Effectiveness of Video Presentation to Students' Learning. *International Journal of Nursing Science*, *5*(2), 81–86.

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science*, *6*(1), 42. https://doi.org/10.1186/1748-5908-6-42

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119.

Mirriahi, N., Joksimović, S., Gašević, D., & Dawson, S. (2018). Effects of instructional conditions and experience on student reflection: A video annotation study. *Higher Education Research & Development*, *37*(6), 1245–1259. https://doi.org/10.1080/07294360.2018.1473845

Mirriahi, N., Jovanovic, J., Dawson, S., Gašević, D., & Pardo, A. (2018). Identifying engagement patterns with video annotation activities: A case study in professional development. *Australasian Journal of Educational Technology*, *34*(1), 57–72. https://doi.org/10.14742/ajet.3207

Mirriahi, N., Jovanović, J., Lim, L.-A., & Lodge, J. M. (2021). Two sides of the same coin: Video annotations and in-video questions for active learning. *Educational Technology Research and Development*, *69*, 2571–2588. https://doi.org/10.1007/s11423-021-10041-4

Mirriahi, N., Liaqat, D., Dawson, S., & Gašević, D. (2016). Uncovering student learning profiles with a video annotation tool: Reflective learning with and without instructional norms. *Educational Technology Research and Development*, *64*(6), 1083–1106. https://doi.org/10.1007/s11423-016-9449-2

Mitrovic, A., Dimitrova, V., Lau, L., Weerasinghe, A., & Mathews, M. (2017). Supporting Constructive Video-Based Learning: Requirements Elicitation from Exploratory Studies. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (pp. 224–237). Springer International Publishing.

Mitrovic, A., Dimitrova, V., Weerasinghe, A., & Lau, L. (2016). Reflective Experiential Learning: Using Active Video Watching for Soft Skills Training. *Proceedings of the 24th International Conference on Computers in Education*, 192–201.

Mitrovic, A., Gordon, M., Piotrkowicz, A., & Dimitrova, V. (2019). Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching. *The 20th International Conference on Artificial Intelligence in Education*, 320–332.

Mitrovic, A., Gostomski, P., Herritsch, A., & Dimitrova, V. (2017). Improving Presentation Skills of First-Year Engineering Students using Active Video Watching. In N. Huda, D. Inglis, N. Tse, & G. Town (Eds.), *Proceedings of the 28th Annual Conference of the Australasian Association for Engineering Education* (pp. 809–816).

Mitrovic, A., & Martin, B. (2007). Evaluating the Effect of Open Student Models on Self-Assessment. *International Journal of Artificial Intelligence in Education, Special Issue on Open Learner Modeling*, *17*(2), 121–144.

Mohammadhassan, N., & Mitrovic, A. (2021a). Investigating Engagement and Learning Differences between Native and EFL students in Active Video Watching. In Ma. M. T. Rodrigo, S. Iyer, & A. Mitrovic (Eds.), *Proceedings of the 29th International Conference on Computers in Education* (Vol. 1, pp. 1–10). Asia-Pacific Society for Computers in Education.

Mohammadhassan, N., & Mitrovic, A. (2021b). Providing Personalized Nudges for Improving Comments Quality in Active Video Watching. *Companion Proceedings of the 11th International Conference on Learning Analytics & Knowledge LAK 2021*, 145.

Mohammadhassan, N., & Mitrovic, A. (2022a). Discovering Differences in Learning Behaviours During Active Video Watching Using Epistemic Network Analysis. In B. Wasson & S. Zörgő (Eds.), *Advances in Quantitative Ethnography* (pp. 362–377). Springer International Publishing.

Mohammadhassan, N., & Mitrovic, A. (2022b). *Investigating the Effectiveness of Visual Learning Analytics in Active Video Watching*. The 23rd International Conference on Artificial Intelligence in Education (AIED 2022) [in print].

Mohammadhassan, N., Mitrovic, A., & Neshatian, K. (2022). Investigating the effect of nudges for improving comment quality in active video watching. *Computers & Education*, *176*, 104340. https://doi.org/10.1016/j.compedu.2021.104340

Mohammadhassan, N., Mitrovic, A., Neshatian, K., & Dunn, J. (2020). Automatic Assessment of Comment Quality in Active Video Watching. *The 28th International Conference on Computers in Education (ICCE 2020)*, 1–10.

Mohammed, A., & Dimitrova, V. (2020). Characterising Video Segments to Support Learning. *The 28th International Conference on Computers in Education (ICCE2020)*, 11–20.

Muldner, K., Wixon, M., Rai, D., Burleson, W., Woolf, B., & Arroyo, I. (2015). Exploring the Impact of a Learning Dashboard on Student Affect. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial Intelligence in Education* (pp. 307–317). Springer International Publishing. https://doi.org/10.1007/978-3-319-19773-9_31

Münscher, R., Vetter, M., & Scheuerle, T. (2016). A Review and Taxonomy of Choice Architecture Techniques. *Journal of Behavioral Decision Making*, *29*(5), 511–524. https://doi.org/10.1002/bdm.1897

Musa, J., Mitrovic, A., Galster, M., & Malinen, S. (2021). Improving Face-to-Face Communication Skills using Active Video Watching. *29th International Conference on Computers in Education Conference (ICCE2021)*, *2*, 683–685.

Okumoto, H., Yoshida, M., Umemura, K., & Ichikawa, Y. (2018). Response Collector: A Video Learning System for Flipped Classrooms. *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, 176–181. https://doi.org/10.1109/ICAICTA.2018.8541338

Paquette, L., Grant, T., Zhang, Y., Biswas, G., & Baker, R. (2021). Using Epistemic Networks to Analyze Self-regulated Learning in an Open-Ended Problem-Solving Environment. In A. R. Ruis & S. B. Lee (Eds.), *Advances in Quantitative Ethnography* (pp. 185–201). Springer International Publishing. https://doi.org/10.1007/978-3-030-67788-6_13

Pardo, A., Mirriahi, N., Dawson, S., Zhao, Y., Zhao, A., & Gašević, D. (2015). Identifying learning strategies associated with active use of video annotation software. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 255–259. https://doi.org/10.1145/2723576.2723611

Pedregosa, F. (2015). *Feature extraction and supervised learning on fMRI: from practice to theory. Doctoral dissertation*.

Pennebaker, J., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin. https://doi.org/10.15781/T29G6Z

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http://www.aclweb.org/anthology/D14-1162

Pintrich, P. R., & de Groot, E. V. (1990). Motivational and Self-Regulated Learning Components of Classroom Academic Performance. *Journal of Educational Psychology*, *82*, 33–40. https://doi.org/10.1037//0022-0663.82.1.33

Popping, R. (2000). *Computer-assisted Text Analysis* (p. 97). SAGE Publications, Ltd. https://doi.org/10.4135/9781849208741.n6

Prenner, N., Klünder, J., & Schneider, K. (2018). Making Meeting Success Measurable by Participants' Feedback. *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering*, 25–31. https://doi.org/10.1145/3194932.3194933

Prevost, L. B., Haudek, K. C., Henry, E. N., Berry, M. C., & Urban-Lurain, M. (2013). Automated text analysis facilitates using written formative assessments for just-in-time teaching in large enrollment courses. *ASEE Annual Conference and Exposition, Conference Proceedings*, 23–26.

Rennie, J., & Srebro, N. (2005). Loss functions for preference levels: Regression with discrete ordered labels. *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, 180–186.

Rice, P., Beeson, P., & Blackmore-Wright, J. (2019). Evaluating the Impact of a Quiz Question within an Educational Video. *TechTrends*, *63*(5), 522–532. https://doi.org/10.1007/s11528-019-00374-6

Risko, E. F., T. Foulsham, S. Dawson, & A. Kingstone. (2013). The Collaborative Lecture Annotation System (CLAS): A New TOOL for Distributed Learning. *IEEE Transactions on Learning Technologies*, *6*(1), 4–13. https://doi.org/10.1109/TLT.2012.15

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rovai, A. P., Wighting, M. J., Baker, J. D., & Grooms, L. D. (2009). Development of an instrument to measure perceived cognitive, affective, and psychomotor learning in traditional and virtual classroom higher education settings. *The Internet and Higher Education*, *12*(1), 7–13. https://doi.org/10.1016/j.iheduc.2008.10.002

Ruiz, J. S., Díaz, H. J. P., Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Kloos, C. D. (2014). Towards the Development of a Learning Analytics Extension in Open EdX. *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality*, 299–306. https://doi.org/10.1145/2669711.2669914

Ruiz, S., Charleer, S., Urretavizcaya, M., Klerkx, J., Fernández-Castro, I., & Duval, E. (2016). Supporting Learning by Considering Emotions: Tracking and Visualization a Case Study. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 254–263. https://doi.org/10.1145/2883851.2883888

Rwitajit, M., Akçapınar, A., Akçapinar, G., Flanagan, B., & Ogata, H. (2019). LAViEW: Learning Analytics Dashboard Towards Evidence-based Education. *The 9th International Conference on Learning Analytics Knowledge (LAK19)*, 68–73.

Saint, J., Gašević, D., Matcha, W., Uzir, N. A., & Pardo, A. (2020). Combining Analytic Methods to Unlock Sequential and Temporal Patterns of Self-Regulated Learning. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 402–411. https://doi.org/10.1145/3375462.3375487

Santos, J. L., Govaerts, S., Verbert, K., & Duval, E. (2012). Goal-oriented visualizations of activity tracking: A case study with engineering students. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 143–152. https://doi.org/10.1145/2330601.2330639

Santos, J. L., Klerkx, J., Duval, E., Gago, D., & Rodríguez, L. (2014). Success, Activity and Drop-Outs in MOOCs an Exploratory Study on the UNED COMA Courses.

Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, 98–102. https://doi.org/10.1145/2567574.2567627

Santucci, V., Sbardella, T., Biscarini, C., Spina, S., & Grego Bolli, G. (2019). Soft skills with learning technologies: The project at the University for Foreigners of Perugia. *Proceedings of the 13th International Technology, Education and Development Conference (INTED 2019)*, 3700–3708.

Schulten, C., Manske, S., Langner-Thiele, A., & Hoppe, H. U. (2020). Bridging Over from Learning Videos to Learning Resources Through Automatic Keyword Extraction. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 382–386). Springer International Publishing.

Scianna, J., Gagnon, D., & Knowles, B. (2021). Counting the Game: Visualizing Changes in Play by Incorporating Game Events. In A. R. Ruis & S. B. Lee (Eds.), *Advances in Quantitative Ethnography* (pp. 218–231). Springer International Publishing. https://doi.org/10.1007/978-3-030-67788-6_15

Sedelmaier, Y., & Landes, D. (2018). Practicing Soft Skills in Software Engineering: A Project-Based Didactical Approach. In Information Resources Management Association (Ed.), *Computer Systems and Software Engineering: Concepts, Methodologies, Tools, and Applications* (pp. 232–252). IGI Global. https://doi.org/10.4018/978-1-5225-3923-0.ch011

Sedrakyan, G., Malmberg, J., Verbert, K., Järvelä, S., & Kirschner, P. A. (2020). Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior*, *107*, 105512. https://doi.org/10.1016/j.chb.2018.05.004

Seo, K., Dodson, S., Harandi, N. M., Roberson, N., Fels, S., & Roll, I. (2021). Active learning with online video: The impact of learning context on engagement. *Computers & Education*, *165*, 104132. https://doi.org/10.1016/j.compedu.2021.104132

Shabaninejad, S., Khosravi, H., Leemans, S. J. J., Sadiq, S., & Indulska, M. (2020). Recommending Insightful Drill-Downs Based on Learning Processes for Learning Analytics Dashboards. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 486–499). Springer International Publishing. https://doi.org/10.1007/978-3-030-52237-7_39

Shaffer, D. (2017). *Quantitative ethnography*. Madison: Cathcart Press.

Shaffer, D., Collier, W., & Ruis, A. R. (2016). A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social, and Interaction Data. *Journal of Learning Analytics*, *3*(3), 9–45. https://doi.org/10.18608/jla.2016.33.3

Shaffer, D., & Ruis, A. (2017). Epistemic Network Analysis: A Worked Example of Theory-Based Learning Analytics. In C. Lang, G. Siemens, A. F. Wise, & D. Gaševic (Eds.), *The Handbook of Learning Analytics* (1st ed., pp. 175–187). Society for Learning Analytics Research (SoLAR). http://solaresearch.org/hla-17/hla17-chapter1

Shelton, C. C., Warren, A. E., & Archambault, L. M. (2016). Exploring the Use of Interactive Digital Storytelling Video: Promoting Student Engagement and Learning in a University Hybrid Course. *TechTrends*, *60*(5), 465–474. https://doi.org/10.1007/s11528-016-0082-z

Sherin, M. G., & Es, E. A. van. (2009). Effects of Video Club Participation on Teachers' Professional Vision. *Journal of Teacher Education*, *60*(1), 20–37.

Shin, H., Ko, E.-Y., Williams, J. J., & Kim, J. (2018). Understanding the Effect of In-Video Prompting on Learners and Instructors. *Proceedings of the 2018 CHI Conference on*

Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3173574.3173893

Sibthorp, J. (2003). Learning transferable skills through adventure education: The role of an authentic process. *Journal of Adventure Education and Outdoor Learning*, *3*(2), 145–157. https://doi.org/10.1080/14729670385200331

Snelson, C., Rice, K., & Wyzard, C. (2012). Research priorities for YouTube and video-sharing technologies: A Delphi study. *British Journal of Educational Technology*, *43*(1), 119–129. https://doi.org/10.1111/j.1467-8535.2010.01168.x

Srivastava, N., Velloso, E., Lodge, J. M., Erfani, S., & Bailey, J. (2019). Continuous Evaluation of Video Lectures from Real-Time Difficulty Self-Report. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300816

Steffes, E. M., & Duverger, P. (2012). Edutainment with Videos and its Positive Effect on Long Term Memory. *Journal for Advancement of Marketing Education*, *20*(1), 1–10.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing Automated Guidance to Promote Productive Revision of Science Explanations. *International Journal of Artificial Intelligence in Education*, *27*(4), 729–757. https://doi.org/10.1007/s40593-017-0145-0

Taskin, Y., Hecking, T., Hoppe, H. U., Dimitrova, V., & Mitrovic, A. (2019). Characterizing Comment Types and Levels of Engagement in Video-Based Learning as a Basis for Adaptive Nudging. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming Learning with Meaningful Technologies* (pp. 362–376). Springer International Publishing.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* Yale University Press.

Tucker, B. (2012). The flipped classroom. *Education Next*, *12*(1), 82–83.

Ullmann, T. D. (2017). Reflective writing analytics: Empirically determined keywords of written reflection. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 163–167.

Ullmann, T. D. (2019). Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches. *International Journal of Artificial Intelligence in Education*, *29*(2), 217–257. https://doi.org/10.1007/s40593-019-00174-2

Ulrich, F., Helms, N. H., Frandsen, U. P., & Rafn, A. V. (2021). Learning effectiveness of 360° video: Experiences from a controlled experiment in healthcare education. *Interactive Learning Environments*, *29*(1), 98–111. https://doi.org/10.1080/10494820.2019.1579234

UNESCO. (2020, March 24). *COVID-19 Educational Disruption and Response*. UNESCO. https://en.unesco.org/news/covid-19-educational-disruption-and-response

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., & Klerkx, J. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing*, *18*(6), 1499–1514. https://doi.org/10.1007/s00779-013-0751-2

Vivian, R., Tarmazdi, H., Falkner, K., Falkner, N., & Szabo, C. (2015). *The Development of A Dashboard Tool for Visualising Online Teamwork Discussions*. 380–388. https://doi.org/10.1109/ICSE.2015.170

Wachtler, J., Hubmann, M., Zöhrer, H., & Ebner, M. (2016). An analysis of the use and effect of questions in interactive learning-videos. *Smart Learning Environments*, *3*(1), 13. https://doi.org/10.1186/s40561-016-0033-3

Wang, M., Peng, J., Cheng, B., Zhou, H., & Liu, J. (2011). Knowledge Visualization for Self-Regulated Learning. *Journal of Educational Technology & Society*, *14*(3), 28–42. JSTOR.

Wang, X., Wen, M., & Rosé, C. P. (2016). Towards Triggering Higher-order Thinking Behaviors in MOOCs. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 398–407. https://doi.org/10.1145/2883851.2883964

Wu, B., Hu, Y., Ruis, A. R., & Wang, M. (2019). Analysing computational thinking in collaborative programming: A quantitative ethnography approach. *Journal of Computer Assisted Learning*, *35*(3), 421–434.

Wylie, R., Sheng, M., Mitamura, T., & Koedinger, K. R. (2011). Effects of Adaptive Prompted Self-explanation on Robust Learning of Second Language Grammar. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial Intelligence in Education* (pp. 588–590). Springer Berlin Heidelberg.

Xia, J., & Wilson, D. C. (2018). Instructor Perspectives on Comparative Heatmap Visualizations of Student Engagement with Lecture Video. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 251–256. https://doi.org/10.1145/3159450.3159487

Yarmand, M., Palani, S., & Klemmer, S. (2021). Adjacent Display of Relevant Discussion Helps Resolve Confusion. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1), 1–10. https://doi.org/10.1145/3449217

Yarmand, M., Yoon, D., Dodson, S., Roll, I., & Fels, S. S. (2019). “Can You Believe [1:21]?!”: Content and Time-Based Reference Patterns in Video Comments. *Proceedings of the*

2019 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3290605.3300719

Yoon, M., Hill, J., & Kim, D. (2021). Designing supports for promoting self-regulated learning in the flipped classroom. *Journal of Computing in Higher Education*, *33*, 398–418. https://doi.org/10.1007/s12528-021-09269-z

Yoon, M., Lee, J., & Jo, I.-H. (2021). Video learning analytics: Investigating behavioral patterns and learner clusters in video-based online learning. *The Internet and Higher Education*, *50*, 100806. https://doi.org/10.1016/j.iheduc.2021.100806

Yousef, A. M. F., Chatti, M. A., & Schroeder, U. (2014). The state of video-based learning: A review and future perspectives. *International Journal of Advanced Life Sciences*, *6*(3/4), 122–135.

Zhang, D., Zhou, L., Briggs, R. O., & Nunamaker, J. F. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information & Management*, *43*(1), 15–27. https://doi.org/10.1016/j.im.2005.01.004

Zhang, S., Meng, X., Liu, C., Zhao, S., Sehgal, V., & Fjeld, M. (2019). ScaffoMapping: Assisting Concept Mapping for Video Learners. In D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, & P. Zaphiris (Eds.), *Human-Computer Interaction – INTERACT 2019* (pp. 314–328). Springer International Publishing.

Zhou, J., & Bhat, S. (2021). Modeling Consistency Using Engagement Patterns in Online Courses. *LAK21: 11th International Learning Analytics and Knowledge Conference*, 226–236. https://doi.org/10.1145/3448139.3448161

Zimmerman, B. J. (2000). Chapter 2 - Attaining Self-Regulation: A Social Cognitive Perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

Appendix A – LIWC Features

(adopted from (Pennebaker et al., 2015))

Category	Subcategory	Abbreviation	Examples	Words incategory
Word count		WC		
Summary Variables	Analytical thinking	Analytic		
	Clout	Clout		
	Authentic	Authentic		
	Emotional tone	Tone		
Words/sentence		WPS		
Words > 6 letters		Sixltr		
Dictionary words		Dic		
Linguistic Dimensions	Total function words	funct	it, to, no, very	491
	Total pronouns	pronoun	I, them, itself	153
	Personal pronouns	ppron	I, them, her	93
	1st pers singular	i	I, me, mine	24
	1st pers plural	we	we, us, our	12
	2nd person	you	you, your, thou	30
	3rd pers singular	shehe	she, her, him	17
	3rd pers plural	they	they, their, they'd	11
	Impersonal pronouns	ipron	it, it's, those	59
	Articles	article	a, an, the	3
	Prepositions	prep	to, with, above	74
	Auxiliary verbs	auxverb	am, will, have	141
	Common Adverbs	adverb	very, really	140
	Conjunctions	conj	and, but, whereas	43
	Negations	negate	no, not, never	62
Other Grammar	Common verbs	verb	eat, come, carry	1000
	Common adjectives	adj	free, happy, long	764
	Comparisons	compare	greater, best, after	317
	Interrogatives	interrog	how, when, what	48
	Numbers	number	second, thousand	36
	Quantifiers	quant	few, many, much	77
Psychological Processes	Affective processes	affect	happy, cried	1393
	Positive emotion	posemo	love, nice, sweet	620
	Negative emotion	negemo	hurt, ugly, nasty	744
	Social processes	social	mate, talk, they	756
	Cognitive processes	cogproc	cause, know, ought	797
	Insight	insight	think, know	259
	Causation	cause	because, effect	135
	Discrepancy	discrep	should, would	83
	Tentative	tentat	maybe, perhaps	178
	Certainty	certain	always, never	113
	Differentiation	differ	hasn't, but, else	81
	Perceptual processes	percept	look, heard, feeling	436
	Drives	drives		1103
	Affiliation	affiliation	ally, friend, social	248
	Achievement	achieve	win, success, better	213
	Power	power	superior, bully	518
	Reward	reward	take, prize, benefit	120
	Risk	risk	danger, doubt	103
Time orientations	Past focus	focuspast	ago, did, talked	341
	Present focus	focuspresent	today, is, now	424
	Future focus	focusfuture	may, will, soon	97
Relativity		relativ	area, bend, exit	974
Spoken categories	Informal language	informal		380
	Swear words	swear	damn, shit	131
	Netspeak	netspeak	btw, lol, thx	209
	Assent	assent	agree, OK, yes	36
	Nonfluencies	nonflu	er, hm, umm	19
	Fillers	filler	Imean, youknow	14

Appendix B – Quality Nudges Study (2020) Documents

B.1 Ethical Approval

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

HUMAN ETHICS COMMITTEE

Secretary, Rebecca Robinson
Telephone: +64 03 369 4588, Extn 94588
Email: human-ethics@canterbury.ac.nz

Ref: HEC 2020/12/LR-PS

2 March 2020

Negar Mohammadhassan
Computer Science and Software Engineering
UNIVERSITY OF CANTERBURY

Dear Negar

Thank you for submitting your low risk application to the Human Ethics Committee for the research proposal titled "Improving the Quality of Comments Using Personalised Nudges in Active Video Watching (AVW)".

I am pleased to advise that this application has been reviewed and approved.

Please note that this approval is subject to the incorporation of the amendments you have provided in your email of 27th February 2020.

With best wishes for your project.

Yours sincerely

Dr Dean Sutherland
Chair, Human Ethics Committee

University of Canterbury Private Bag 4800, Christchurch 8140, New Zealand. www.canterbury.ac.nz

F E S

B.2 Information Sheet

Department: Computer Science and Software Engineering
Email: Negar.Mohammadhassan@pg.canterbury.ac.nz
HEC Ref: HEC 2020/12/LR-PS

5/02/2020

Invitation to participate in the Active Video Watching study

I am a PhD student in the Department of Computer Science and Software Engineering at the University of Canterbury. I am conducting a research project that investigates the effectiveness of a system called Active Video Watching (AVW-Space) for improving presentation skills.

I would like to invite you to participate in my study. AVW-Space is a Web-based platform developed by the Intelligent Computer Tutoring Group (ICTG) at the University of Canterbury. If you agree to participate, you can use your UC user code to log into AVW-Space. We have chosen a set of videos that should enhance your knowledge of how to pitch your ideas to a general audience. You will be able to use AVW-Space in your own time. This platform has two phases: 1) Personal Space where you can watch videos and write comments on the videos, and 2) Social Space where you can review comments written by others and rate them.

If you choose to take part in this study, your involvement in this project will be: 1) completing a profile survey, 2) using AVW-Space and 3) completing a survey about your experience with AVW-Space. The overall time required for these activities is approximately 2-3 hours.

Participation is voluntary and you can give your consent for participating in the study as the initial step of the profile survey. You have the right to withdraw at any stage without penalty. You may ask for your raw data to be returned to you or destroyed at any point. At the end of the study, there will be a lucky draw including all students who completed the study. The prizes are two vouchers worth $100 each.

This study is under the supervision of Professor Tanja Mitrovic (Tanja.Mitrovic@canterbury.ac.nz), Kourosh Neshatian (Kourosh.Neshatian@canterbury.ac.nz) and Jonathan Dunn (Jonathan.Dunn@canterbury.ac.nz). My main supervisor, Professor Tanja Mitrovic, will be pleased to discuss any concerns you may have about participation in the project.

The results of the study may be published, but you may be assured of the complete confidentiality: you will not be identified in any reports or publications. Your responses to the surveys and your interaction with AVW-Space will be stored securely on the University of Canterbury servers. This data will be accessible only to Negar Mohammadhassan, Professor Tanja Mitrovic and the research team working on AVW-Space (including potential future researchers). All data collected in this study will be destroyed after 10 years since this is a part of my PhD thesis. The data analysis will be published as part of a PhD thesis and academic papers. A thesis is a public document and will be available through the UC Library. You can email me if you would like to receive a summary of the results of the project.

This project has been reviewed and approved by the University of Canterbury Human Ethics Committee, and participants should address any complaints to The Chair, Human Ethics Committee, University of Canterbury, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz).

B.3 Consent Form

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Department: Computer Science and Software Engineering
Email: Negar.Mohammadhassan@pg.canterbury.ac.nz

Improving the quality of comments using personalised nudges in Active Video Watching

I have been given a full explanation of this project and have had the opportunity to ask questions.

I understand what is required of me if I agree to take part in the research.

I understand that participation is voluntary and I may withdraw at any time without penalty. Withdrawal of participation will also include the withdrawal of any information I have provided should this remain practically achievable.

I understand that any information or opinions I provide will be kept confidential to the researcher, her supervisor and the research team, and that any published or reported results will not identify the participants. I understand that a thesis is a public document and will be available through the UC Library.

I understand that all data collected for the study will be kept in locked and secure facilities and/or in password-protected electronic form and will be destroyed after ten years.
I understand the risks associated with taking part and how they will be managed.

I understand that I can contact the researcher Negar Mohammadhassan (negar.mohammadhassan@pg.canterbury.ac.nz) or supervisor Professor Tanja Mitrovic (Tanja.mitrovic@canterbury.ac.nz) for further information. If I have any complaints, I can contact the Chair of the University of Canterbury Human Ethics Committee, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz)

By signing below, I agree to participate in this research project.

Name:_Signed:______Date:__

Email address (for report of findings, if applicable):_

B.4 Questionnaires

B.4.1 Survey 1

Q1. Please enter your UC username (for consent form)

The following questions ask about your profile.

Q2. What is your age?

- ❍ 18-23 (1)
- ❍ 24-29 (2)
- ❍ 30-35 (3)
- ❍ 36-41 (4)
- ❍ 42-47 (5)
- ❍ 48+ (6)

Q3. What is your gender?

- ❍ Male (1)
- ❍ Female (2)
- ❍ Other (3)

Q4. What is your first language?

Q5. How much formal training have you had on presentation skills?

- ❍ No training (1)
- ❍ Some training (2)
- ❍ Quite a Bit (3)
- ❍ A lot (4)
- ❍ Extensive training (5)

Answer if “No training” is not selected for Q5.

Q6. Select the type(s) of training about presenting you have had:

- ❑ Training at high school (1)
- ❑ Training at University (2)
- ❑ Practice with feedback (3)
- ❑ Professional development training (4)
- ❑ Other (please specify) (5) ____________________

Q7. Please indicate your experience in giving presentations:

- ❍ Not experienced (1)
- ❍ A little (2)
- ❍ Medium level (3)
- ❍ A lot (4)
- ❍ Highly experienced (5)

Answer if “Not experienced” is not selected for Q7.

Q8. Please specify the type(s) of presentations you have given:

- ❑ project presentation (1)
- ❑ course work presentation (2)
- ❑ seminar (3)
- ❑ conference presentation (4)
- ❑ pitching an idea (5)
- ❑ outreach presentation (6)
- ❑ presentation for a general audience (7)
- ❑ other (please specify) (8) ____________________

Q9. How often do you watch YouTube?

- ❍ Never (1)
- ❍ Occasionally (2)
- ❍ Once a month (3)
- ❍ Every week (4)
- ❍ Every day (5)

Q10. How often do you use YouTube for learning?

- ❍ Never (1)
- ❍ Occasionally (2)
- ❍ Once a month (3)
- ❍ Every week (4)
- ❍ Every day (5)

Q11. The following questions ask about your motivation for and attitudes about studying and should not take more than a few minutes. Remember, there are no right or wrong answers; just answer as accurately as possible. Please enter your rankings for the following statements. [The following questions all have the Likert scale 1 (Not at all true of me) to 7 (Very true of me)]

- I prefer training material that really challenges me so I can learn new things.
- I think I will be able to use what I learn in future.
- I believe I will receive excellent grades for my performance.
- I am certain I can understand the most difficult training material.
- Getting high recognition for my performance is the most satisfying thing for me right now.
- It is my own fault if I do not learn the material in training courses.
- The most important thing for me right now is getting good grades.
- I am confident I can understand the basic concepts presented in the training material.
- If I can, I want to get better grades than most of the other students.
- I have a great deal of control over my performance in my courses/studies.
- I prefer course material that arouses my curiosity, even if it is difficult to learn.
- If I try hard enough, I will understand training materials.
- I am confident I can do an excellent job on the assessment of my performance.
- I expect to do well in my studies.
- The most satisfying thing for me is to understand the content as thoroughly as possible.
- When I have the opportunity, I choose assignments that I can learn from even if they do not guarantee a good grade.
- I like the subject matter of my courses/studies.
- Understanding the subject matter is very important to me.
- I am certain I can master the skills taught.
- I want to do well in my studies because it is important to show my ability to my family, friends, employer and others.

- During training sessions, I often miss important points because I am thinking of other things.
- The more effort I put into my studies, the better I do.
- When studying, I make up questions to help focus.
- I often feel lazy or bored when studying that I quit before I finish what I planned to do.
- I practice saying the material to myself over and over.
- When I become confused about something, I go back and try to figure it out.
- If training materials are difficult to understand, I change the way I read them.
- I go through the materials over and over.
- I work hard to do well in my studies even if I do not like what I am doing.
- I make simple charts, diagrams or tables to help me organize the material.
- I pull together information from different sources, such as lectures, online material and discussions.
- Before I study the new material thoroughly, I often skim it to see how it is organized.
- I ask myself questions to make sure I understand the material.
- I often find that I have been reading but do not know what it was all about.
- I memorize the keywords to remind myself of important concepts.
- When an assignment is difficult, I give up or only attempt the easy parts.
- I try to think through a topic and decide what I am supposed to learn from it rather than just reading it over.
- I try to relate ideas in one course to those in other courses whenever possible.
- I go over my notes and make an outline of important concepts.
- I try to relate the material to what I already know.
- I write brief summaries of the main ideas from the material.
- I make lists of important items for my courses and memorize the lists.
- Even when training materials are dull and uninteresting, I manage to keep working until I finish.
- I try to determine which concepts I do not understand well.
- I set goals for myself in order to direct my activities in each study period.
- I am able to evaluate for myself how I make progress at learning.

The following questions ask about giving presentations.

Q12. [You have max 60 seconds to answer] Write all words/phrases (one per line) that you associate with STRUCTURE (as related to giving presentations)

Q13. [You have max 60 seconds to answer] Write all words/phrases (one per line) that you associate with DELIVERY & SPEECH (as related to giving presentations)

Q14. [You have max 60 seconds to answer] Write all words/phrases (one per line) that you associate with VISUAL AIDS (as related to giving presentations)

B.4.2 Survey 2

Q1. Please enter your user code

Note that all your answers are kept in an anonymous way.

This questionnaire includes **two parts**:

1) terms related to giving presentations,
2) feedback on AVW-Space.

The following questions ask about **giving presentations**.

Q2. [You have **max 60 seconds** to answer] Write **all words/phrases** (one per line) that you associate with **STRUCTURE (as related to giving presentations)**

Q3. [You have **max 60 seconds** to answer] Write **all words/phrases** (one per line) that you associate with **DELIVERY & SPEECH (as related to giving presentations)**

Q4. [You have **max 60 seconds** to answer] Write **all words/phrases** (one per line) that you associate with **VISUAL AIDS (as related to giving presentations)**

The following questions gather your **feedback on AVW-Space**. [The following questions have a Likert scale from 1 (Very easy) to 20 (Very Hard)]

Q5. **MENTAL DEMAND - Writing comments**

How **mentally demanding** was to **write comments on videos** in AVW-Space? For example, how much mental and perceptual activity was required - thinking, deciding, remembering, looking, searching?

Q6. **EFFORT - Writing comments**

How **hard** did you have to work (mentally and physically) to **write comments on videos** in AVW-Space?

Q7. **FRUSTRATION - Writing comments**

How **discouraged, irritated, stressed and annoyed** did you feel while **writing comments on videos** in AVW-Space?

Q8. **PERFORMANCE - Writing comments**

How successful do you think you were to identify useful points about presentation skills when commenting on videos in AVW-Space?

Q9. Based on your use of AVW-Space, what would be the usefulness of **pausing a video to write a comment**?

Q10. Based on your use of AVW-Space, what would be the usefulness of **asking you to indicate what the comments refer to** (e.g., **for tutorials:** '*I am rather good at this*', '*I did/saw this in the past*', '*I did not realise I was not doing this*', '*I like this point*'; **for examples:** '*Structure*', '*Delivery*', *'Speech'*, '*Visual Aids*'.)

Q11. **MENTAL DEMAND - Rating comments**

How **mentally demanding** was to **review and rate comments on videos** in AVW-Space?

For example, how much mental and perceptual activity was required - thinking, deciding, remembering, looking, searching?

Q12. **EFFORT - Rating comments**

How **hard** did you have to work (mentally and physically) to **review and rate comments on videos** in AVW-Space?

Q13. **FRUSTRATION - Rating comments**

How **discouraged, irritated, stressed and annoyed** did you feel while **reviewing and rating the comments on videos** in AVW-Space?

Q14. **PERFORMANCE - Rating comments**

How **successful** do you think you were to **identify useful points about presentation skills** when **reviewing and rating of comments** made by others in AVW-Space?

Q15. The AVW-Space system is aimed at **informal learning of soft skills** (e.g., giving presentations, advising, negotiating, managing teams) using selected videos.

The questions below ask how you **perceive the usefulness** of AVW-Space for informal learning of soft skills. [The following questions are in the Likert scale from 1(extremely likely) to 7 (extremely unlikely)]

-I think I would like to use AVW-Space frequently.
-I would recommend AVW-Space to my friends.
-Using AVW-Space would enable me to improve my soft skills quickly.
-Using AVW-Space would improve my performance, considering the development of soft skills.
-Using AVW-Space would enhance my effectiveness when developing soft skills.
-I would find AVW-Space useful in my studies/job.
-I would find AVW-Space easy to do what I want it to do.
-My interaction with AVW-Space would be clear and understandable.
-I would find AVW-Space easy to use.
-If I am provided the opportunity, I would continue to use AVW-Space for informal learning.

Q17. AVW-Space provides **two features for active video watching**:

- visualisations of previous comments
- hints on active video watching

Based on your use of AVW-Space, what would be the usefulness of **provided visualisations** (i.e., the dots below the video representing others comments and your comments).

Q18. Based on your use of AVW-Space, what would be the usefulness of **the hints you received** (i.e., additional instructions that appeared to the right of the video)?

Q19. **In the second phase of the study, you experienced two additional features of AVW-Space**:

- reviewing the comments on the videos made by other users of AVW-Space;
- rating the comments of other users;

Based on your use of the AVW-Space, what would be the usefulness o**f reviewing the comments on the videos made by other people?**

Q20. Based on your use of AVW-Space, what would be the usefulness of **rating the comments of other people**? For example :

'This is useful for me'
'I hadn't thought of this'
'I did not notice this'
'I like this point'
'I do not agree with this'

Q21. What do you think is **most exciting** about AVW-Space?

Q22. What do you think is **most disappointing** about AVW-Space?

Appendix C – Visualisations Study (2021) Documents

C.1 Ethical Approval

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

HUMAN ETHICS COMMITTEE

Secretary, Rebecca Robinson
Telephone: +64 03 369 4588, Extn 94588
Email: human-ethics@canterbury.ac.nz

Ref: HEC 2020/12/LR-PS Amendment 1

7 April 2021

Negar Mohammadhassan
Computer Science and Software Engineering
UNIVERSITY OF CANTERBURY

Dear Negar

Thank you for your request for an amendment to your research proposal "Improving the Quality of Comments Using Personalised Nudges in Active Video Watching (AVW)" as outlined in your email dated 30th March 2021.

I am pleased to advise that this request has been considered and approved by the Human Ethics Committee.

Yours sincerely

Dr Dean Sutherland
Chair, Human Ethics Committee

University of Canterbury Private Bag 4800, Christchurch 8140, New Zealand. www.canterbury.ac.nz

F E S

C.2 Information Sheet and Consent Form

Same as the previous study in 2020 (Please refer to B.2 Information Sheet and B.3 Consent Form)

C.3 Questionnaires

Same as the previous study in 2020 (B.4 Questionnaires). Following questions were added to Survey 2 in the 2021 study:

Q23. Based on your use of AVW-Space, what would be the usefulness of **the provided visualisations of your progress** (i.e., the summary of your activities on the main page)?

Q24. Based on your use of AVW-Space, what would be the usefulness of **displaying the hints you received previously and the predicted quality of your comments** (i.e., The  icon in the comments list)?

Q25. Based on your use of AVW-Space, what would be the usefulness of **showing the ratings on your comments** in the reviewing phase?

Appendix D – Generalisability Study (SENG2021) Documents

D.1 Ethical Approval

UC
UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

HUMAN ETHICS COMMITTEE

Secretary, Rebecca Robinson
Telephone: +64 03 369 4588, Extn 94588
Email: human-ethics@canterbury.ac.nz

Ref: HEC 2020/30/LR-PS Amendment 1

24 June 2021

Ja'afaru Musa
Computer Science and Software Engineering
UNIVERSITY OF CANTERBURY

Dear Ja'afaru

Thank you for your request for an amendment to your research proposal "Reflective Experiential Learning: A Novel Approach for Improving Face-to-Face Communication in Software Development Meetings of Inexperienced Software Engineers Using Active Video Watching" as outlined in your emails dated 9th and 15th June 2021.

I am pleased to advise that this request has been considered and approved by the Human Ethics Committee.

Yours sincerely

Dr Dean Sutherland
Chair, Human Ethics Committee

University of Canterbury Private Bag 4800, Christchurch 8140, New Zealand. www.canterbury.ac.nz

F E S

D.2 Information Sheet

Department of Computer Science and Software Engineering
Telephone: Ext. 94269
Email: jaafaru.musa@pg.canterbury.ac.nz
Date: 26/07/2021
HEC Ref: HEC 2020/30/LR-PS

UC UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Improving face-to-face communication in software development meetings using Active Video Watching

I am Jaafaru Musa, a PhD student in the Department of Computer Science and Software Engineering. This research aims to investigate the effectiveness of active video watching in improving face-to-face communication skills during software development project meetings.

You have been approached to take part in this study because you are currently enrolled in SENG202.

If you choose to take part in this study, your involvement in this project will be to 1) fill in a profile survey; 2) watch and make comments on videos; 3) review and rate anonymized comments made by others; 4) record one team meeting; 5) watch and make comments on your team video; 6) review and rate anonymized comments made by your team members, and 7) complete the exit survey. The total time is estimated at 2-3 hours.

Please note that the video recordings of team meetings will only be available to the team members and the SENG202 teaching team (not to the whole class) and will be uploaded to a private video sharing platform.

Participation is voluntary and you have the right to withdraw at any stage without penalty. You may ask for your raw data to be returned to you or destroyed at any point. If you withdraw, I will remove information relating to you.

The results of the project may be published, but you may be assured of the complete confidentiality of data gathered in this investigation: your identity will not be made public. To ensure anonymity and confidentiality, the collected data will be stored using unique identifiers generated for this study. Any identifying information (names and UC user codes) will be kept separate from the collected data. The data will only be accessible by the research team working on AVW-Space, and will be destroyed 10 years after the completion of the project. A thesis is a public document and will be available through the UC Library. Please email me if you would like to receive a copy of the summary of the results of the project.

The project is being carried out as a requirement for PhD in Computer Science by Jaafaru Musa under the supervision of Professor Tanja Mitrovic (tanja.mitrovic@canterbury.ac.nz), Associate Professor Matthias Galster (matthias.galster@canterbury.ac.nz), and Associate Professor Sanna Malinen (sanna.malinen@canterbury.ac.nz). They will be pleased to discuss any concerns you may have about participation in the project.

This project has been reviewed and approved by the University of Canterbury Human Ethics Committee, and participants should address any complaints to The Chair, Human Ethics Committee, University of Canterbury, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz).

D.3 Consent Form

Department of Computer Science and Software Engineering

Telephone: Ext. 94269

Email: jaafaru.musa@pg.canterbury.ac.nz

Improving face-to-face communication in software development meetings using Active Video Watching

Consent Form for Participants

I have been given a full explanation of this project and have had the opportunity to ask questions.

I understand what is required of me if I agree to take part in the research.

I understand that participation is voluntary and I may withdraw at any time without penalty. Withdrawal of participation will also include the withdrawal of any information I have provided should this remain practically achievable.

I understand that any information or opinions I provide will be kept confidential to the research team working on AVW-Space (including future PG students who join the team), and that any published or reported results will not identify the participants. I understand that a thesis is a public document and will be available through the UC Library.

I understand that all data collected for the study will be kept in password protected electronic form and will be destroyed after ten years.

I understand that I can contact Jaafaru Musa (jaafaru.musa@pg.canterbury.ac.nz) or the supervisor Prof. Tanja Mitrovic (tanja.mitrovic@canterbury.ac.nz) for further information. If I have any complaints, I can contact the Chair of the University of Canterbury Human Ethics Committee, Private Bag 4800, Christchurch (human-ethics@canterbury.ac.nz)

By signing below, I agree to participate in this research project.

Name:_Signed:______Date:__

Email address (for report of findings, if applicable):____

D.4 Questionnaires

D.4.1 Survey 1

Q1. Please enter your UC username (for consent form)

The following questions ask about your profile.

Q2. What is your age?

- ❍ 18-23 (1)
- ❍ 24-29 (2)
- ❍ 30-35 (3)
- ❍ 36-41 (4)
- ❍ 42-47 (5)
- ❍ 48+ (6)

Q3. What is your gender?

- ❍ Male (1)
- ❍ Female (2)
- ❍ Other (3)

Q4. What is your first language?

Q5. How much formal training have you had on communication in face-to-face meetings?

- ❍ No training (1)
- ❍ Some training (2)
- ❍ Quite a Bit (3)
- ❍ A lot (4)
- ❍ Extensive training (5)

Q6. Select the type(s) of training on communication in face-to-face meetings you have had:

- ❑ Training at high school (1)
- ❑ Training at University (2)
- ❑ Practice with feedback (3)
- ❑ Professional development training (4)
- ❑ Other (please specify) (5) ____________________

Q7. Over the last year, how frequently would you attend face-to-face formal meetings with more than two people?

- ❍ Never
- ❍ Occasionally
- ❍ Once a month
- ❍ Every week
- ❍ Every day

Q8. Please specify the type(s) of meetings you have had that involved more than two people:

- ❑ Group assignment in high school
- ❑ Group assignment in university
- ❑ Meeting with lecturers
- ❑ As part of an internship
- ❑ Part-time job related to software development
- ❑ Part-time job not related to software development
- ❑ Other (Please specify) ____________

Q9. How much experience do you have working in software development teams outside the university?

- ❍ None
- ❍ Some experience (less than a week)
- ❍ Quite a Bit (a month)
- ❍ A lot (several months)
- ❍ Extensive experience (more than a year)

Q10. How often do you watch YouTube?

- ❍ Never (1)
- ❍ Occasionally (2)
- ❍ Once a month (3)
- ❍ Every week (4)
- ❍ Every day (5)

Q11. How often do you use YouTube for learning?

- ❍ Never (1)
- ❍ Occasionally (2)
- ❍ Once a month (3)
- ❍ Every week (4)
- ❍ Every day (5)

Q12. [You have max 3 minutes to answer] Write all words/phrases (one per line) that you associate with effective communication in software engineering meetings.

Q13. Please rate on the scale 1 (never) to 7 (always), the level that describes your typical behaviour during face-to-face team meetings: [I do this: 1 (never) to 7 (always).]

VERBAL COMMUNICATION

- I express technical ideas clearly, so that every meeting participant can understand.
- I express non-technical ideas clearly, so that every meeting participant can understand.
- I vary language and expression to suit different situations during team meetings.
- I make eye contact with meeting participants during discussions.

Please add a comment below if you find any of these questions difficult to understand or any other comments.

GIVING AND RECEIVING FEEDBACK

- I provide constructive feedback to other meeting participants.
- I am respectful to other meeting participants.
- I am mindful of other meeting participants feelings when providing feedback.
- I get defensive when receiving other meeting participants' negative feedback.
- I receive other meeting participants' feedback as a constructive contribution.

- I use the team's feedback to improve my participation during team meetings.

Please add a comment below if you find any of these questions difficult to understand or any other comments.

ACTIVE LISTENING

- I often begin to talk before the other meeting participants finish talking.
- I begin arguing with the other meeting participants before I have heard their entire idea.
- When I want to say something, I talk about it, even if I interrupt the other meeting participants.
- I listen to the other meeting participants, putting myself in her/his shoes.
- I pay attention to the other meeting participants body language.
- I am aware of my feelings while I'm listening to other meeting participants.
- If I do not understand what another meeting participant said, I seek clarification by asking questions.

Please add a comment below if you find any of these questions difficult to understand or any other comments.

MEETING PARTICIPATION

- I contribute my ideas and suggestions during team meetings.
- When other meeting participants are hesitating to contribute, I encourage them to contribute their ideas and suggestions.
- I express my personal feelings when I agree with other meeting participants.
- I express my personal feelings when I disagree with other meeting participants.
- I encourage other meeting participants to express their personal feelings.
- I check my mobile, emails or notifications during meetings.

Please add a comment below if you find any of these questions difficult to understand or any other comments.

D.4.2 Survey 2

Note that all your answers are kept in an anonymous way, there no right or wrong answers to the questions and you aren't graded on them.

This questionnaire includes three parts:

1. Question about effective communication in software engineering meetings;
2. Self-assessment of face-to-face meeting communication skills;
3. Questions about how much you have learned from AVW-Space

Q1. Please enter your user code:

Q2. [You have max 3 minutes to answer] Write all words/phrases (one per line) that you associate with effective communication in software engineering meetings.

Q3. Please rate on the scale 1 (never) to 7 (always), the level that describes your typical behaviour during face-to-face team meetings: [I do this: 1 (never) to 7 (always).]

VERBAL COMMUNICATION

- I express technical ideas clearly, so that every meeting participant can understand.
- I express non-technical ideas clearly, so that every meeting participant can understand.
- I vary language and expression to suit different situations during team meetings.
- I make eye contact with meeting participants during discussions.

Please add a comment below if you find any of these questions difficult to understand or any other comments.

GIVING AND RECEIVING FEEDBACK

- I provide constructive feedback to other meeting participants.
- I am respectful to other meeting participants.
- I am mindful of other meeting participants feelings when providing feedback.
- I get defensive when receiving other meeting participants' negative feedback.
- I receive other meeting participants' feedback as a constructive contribution.

- I use the team's feedback to improve my participation during team meetings.

Please add a comment below if you find any of these questions difficult to understand or any other comments.

ACTIVE LISTENING

- I often begin to talk before the other meeting participants finish talking.
- I begin arguing with the other meeting participants before I have heard their entire idea.
- When I want to say something, I talk about it, even if I interrupt the other meeting participants.
- I listen to the other meeting participants, putting myself in her/his shoes.
- I pay attention to the other meeting participants body language.
- I am aware of my feelings while I'm listening to other meeting participants.
- If I do not understand what another meeting participant said, I seek clarification by asking questions.

Please add a comment below if you find any of these questions difficult to understand or any other comments.

MEETING PARTICIPATION

- I contribute my ideas and suggestions during team meetings.
- When other meeting participants are hesitating to contribute, I encourage them to contribute their ideas and suggestions.
- I express my personal feelings when I agree with other meeting participants.
- I express my personal feelings when I disagree with other meeting participants.
- I encourage other meeting participants to express their personal feelings.
- I check my mobile, emails or notifications during meetings.

Please add a comment below if you find any of these questions difficult to understand or any other comments.

Q4. The following questions ask you to estimate how much you have learned from AVW-Space.

Please rate, on the scale of 1 (strongly disagree) to 7 (strongly agree), to what extent do you agree with each statement, where lower numbers reflect less agreement and higher numbers reflect more agreement.

- I can summarize what I have learnt in AVW-Space for someone who has not learned from AVW-Space.
- I am able to use the effective meeting participation concepts I learnt in AVW-Space in my future meetings.
- I have changed my attitudes about effective meeting participation as a result of AVW-Space.
- I can assess the quality of face-to-face communication in the example videos used in AVW-Space.
- I feel more confident in my face-to-face communication skills in meetings as a result of AVW-Space.
- I have not expanded my knowledge of effective meeting participation concepts as a result of AVW-Space.
- I can demonstrate to others the effective meeting participation concepts I learnt in AVW-Space.
- I feel that I am a more effective meeting participant as a result of AVW-Space