A Proposed Methodology For Investigating Chatbot Effects In Peer Review

by

Michael Pin-Chuan Lin

M.B.A, National Formosa University, 2011

B.B.A., National Penghu University of Science and Technology, 2009

Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in the

Educational Technology and Learning Design Program

Faculty of Education

SIMON FRASER UNIVERSITY

Summer 2020

Approval

Name:	**Michael Pin-Chuan Lin**
Degree:	**Doctor of Philosophy**
Title:	**A Proposed Methodology For Investigating Chatbot Effects In Peer Review**
Examining Committee:	**Chair:** Kevin O'Neill Associate Professor
	Philip H. Winne Senior Supervisor Professor
	Fred Popowich Supervisor Professor School of Computing Science
	Engida Gebre Internal Examiner Assistant Professor
	Vivekanandan Kumar External Examiner Professor School of Computing and Information Systems Athabasca University
Date Defended/Approved:	July 24th, 2020

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Teaching academic writing skills consumes a lot of time for teachers. One way to save some of this time and support students' development of writing skills is to supplement teacher-student interaction with a chatbot. I developed such a chatbot, DD, to help post-secondary writers develop a thesis statement for an argumentative essay and to improve their feedback when in the role of a peer reviewer of classmates' draft essays. The study analyzes student-chatbot interactions in a lower division course as background for developing methodological procedures that examine students' engagement patterns with a chatbot. Analyses of student-chatbot data reveal students participating in this study tended to be overconfident about their learning. Furthermore, students reported a positive experience when they conversed with the chatbot. Several pedagogical implications for chatbot-guided writing instructions and uses of learning technology are addressed.

Keywords: Chatbot; Educational Technology; Content Analysis

To DD, Queen, & My Family

DD、Queen、我的家庭

Acknowledgements

First, I am so grateful to have Dr. Philip Winne as my senior supervisor during my Ph.D. journey. You have always supported and guided me through many challenges and helped me to become a better researcher and educator. I cannot express my gratitude enough. I want to thank Dr. Fred Popowich for advising during my study and providing critical and insightful feedback to this dissertation. I would also want to thank my dissertation committee members' feedback, Dr. Kevin O'Neill, Dr. Engida Gebre, and Dr. Vive Kumar for making this dissertation better.

A big thank you to my friend, Daniel Chang, during my Ph.D. studies. You provide all the support, and without your help, I cannot finish this research. It is my pleasure to work with my labmates in the EdPsych lab at Simon Fraser University: John Nesbit, Zahia Marzouk, Mladen Rakovic, Kenny Teng, Donya Samadi, Alexandra Patzak, Jovita Vitasek, Azar Savoji, Aaron Black, Nashwa Nashaat Sobhy, and Jason Stewart-Alonso. Thank you for all the inspiration and support in research.

Most importantly, I cannot complete this dissertation without my family's support. DD, thank you for always being with me and for being the model of the chatbot in my dissertation. Queen, although you are not in Canada due to your health problem, your spirit has always been with me, and I can feel you are with me here. Dad and mom, the gratitude cannot be expressed through words and thanks for all the support, encouraging me during my study, and chatting with me when I am stressed. I am so proud to be the first doctor in our big family.

Table of Contents

List of Tables

List of Figures

Chapter 1.

Introduction

Applications of artificial intelligence (AI) have been forecast for education for many years. Two examples that have been realized, although to different extents, are intelligent tutoring systems (ITS) and chatbots. ITS generally use a model to tailor student learning and to adapt content or style of instructions (Kerly, Hall, & Bull, 2007; Murray, 1999). Ma, Adesope, Nesbit, and Liu's (2014) meta-analysis examined a relatively extensive set of studies investigating whether and how an ITS can promote cognitive, motivational or metacognitive knowledge of an individual student by modelling their psychological states as a basis for providing individualized instruction. However, Murray (1999) pointed out some limitations of ITS, including students' perceptions of low fidelity and limited instructional values, incomplete student modelling, interactivity (learning by reading instead of learning by doing), and insufficiently building students' problem-solving skills. A chatbot is programmed to supplement an ITS with diverse real-time dialogue during learning may more satisfyingly stimulate thinking processes (Kerly et al., 2007; Song, Oh, & Rice, 2017).

Jain, Kumar, Kota, and Patel (2018) defined chatbots as "text-based, turn-based, task-fulfilling programs, embedded within existing platforms" (p. 904). Chatbots are designed to stimulate human-like conversations, answer questions, support users, or tutor learners (Abbasi & Kazi, 2014; Wang & Petrina, 2013). Studies (Clarizia et al., 2018; Mekni, Baani, & Sulieman, 2020) pointed out two potential benefits of chatbots in educational contexts: chatbots can (a) create a one-to-one interactive learning opportunity for students, and (b) enhance student learning by analyzing and tracking their behaviour and improvement. Furthermore, Reiners, Wood, and Bastiaens (2014) expressed that chatbots provide students with an authentic learning environment which could improve student engagement with learning materials. However, only a few studies charted the design of a chatbot and a plan for evaluating student-chatbot interactions from a perspective of instructional design (e.g., Fryer & Carpenter, 2006; Kerly et al., 2007; Wang & Petrina, 2013). Specifically, how students engage with a chatbot during learning activities is underexplored (Abbasi & Kazi, 2014; Clarizia, Colace, Lombardi,

Pascale, & Santaniello, 2018; Fryer, Ainley, Thompson, Gibson, & Sherlock, 2017; Goda, Yamada, Matsukawa, Hata, & Yasunami, 2014; Kerly et al., 2007).

In research about chatbots, evaluation of performance has mainly been judged by its capabilities to manage novel input from users and users' satisfaction with the chatbot. Regarding managing novel user input, a variety of methods have been studied including keyword/string-matching techniques (Abbasi & Kazi, 2014; Ahmad, Baharum, Hamid, & Zainal, 2020; Clarizia et al., 2018; Sarosa, Kusumawardani, Suyono, & Wijaya, 2020; Weizenbaum 1966), machine learning (Mekni et al., 2020; Ranavare & Kamath, 2020; Serban et al., 2017; Sinha, Basak, Dey, & Mondal, 2020), and efficiency of conversational exchange (Pham, Pham, Nguyen, Nguyen, & Cao, 2018). Users satisfaction with a chatbot has been assessed based on perceptions about how well a chatbot helped solved problem(s) and its ease of use (Clarizia et al., 2018; Kerly et al., 2007), its contributions to a conversation/task (Goda et al., 2014; Heller, Proctor, Mah, Jewell, & Cheung, 2005; Song et al., 2017; Wang & Petrina, 2013), how interesting or fun it was to work with (Fryer & Carpenter, 2006), and students interests in task/course long-term development (Fryer et al., 2017). These methods do not, however, offer clear guidance about how to tune a chatbot's design.

In this dissertation, I focus on post-secondary learners' interactions with a chatbot custom-designed to guide the development of an argumentative essay. A successful chatbot supporting this writing task could supplement writing instruction, particularly in large classes where instructors may not available at the time when students need help (Song et al., 2017; Xu & Wang, 2006). The study was implemented in a first-year educational psychology course. One of the graded components of the course was to outline an argumentative essay. Each student's outline was peer reviewed by another student. In prior literature, peer review activities have been noted to help students become better reviewers and better writers because reviewing another student's drafts stimulates the reviewer's critical thinking and evaluation of their own drafts (e.g., Cho & Schunn, 2007; Cho & Cho, 2011). However, peer review is a complex task and students often have difficulties giving useful feedback (Falchikov, 1995; Macdonald, 2001). Some research investigated ways to support students to develop better reviews, including training (Min, 2005), peer review guidance (Cho & Schunn, 2007), or an online support system (Kulkarni, Kotturi, Bernstein, & Klemmer, 2016). Additionally, Sluijsmans, Brand-Gruwel, and van Merriënboer (2002) suggested a

training program for peer review may be promising as means to build students' skills for "judging the performance of a peer reflecting upon and identifying the strengths and weaknesses in a peer's product and writing an assessment report; providing feedback for future learning—giving constructive feedback about the product of a peer" (p. 444). This idea brings up questions about how to design a chatbot from the perspective of principles of instructional design so the chatbot can support students to become better reviewers of peers' draft essays.

To fill the aforementioned research gaps, in this dissertation study I designed a chatbot to improve student reviewers' feedback on a peer's outline of an argumentative essay. Developing good thesis statements is important. Recent findings reported by Chang (2020)' indicate qualities of thesis statements are good predictors of the quality of an essay's introduction and of the overall essay. The chatbot focused on simplifying peer review instructions and enhancing student learning about how to give effective peer feedback. The study analyzes student-to-chatbot interactions to develop methodological guidelines for examining students' engagement patterns with this chatbot. Eight possible interactional patterns with the chatbot were identified and analyzed qualitatively. Deeply examining these patterns provides guidance for developing chatbots to be used in educational settings. Specifically, this dissertation proposes and illustrates a methodology that provides insights into episodes where students work with the chatbot.

Chapter 2.

Theoretical Background and Literature Review

2.1. Constructivism

One foundation stone for this dissertation study is constructivism. Constructivism originated, in part, from Piaget's theory of cognitive development. A constructivist view of learning focuses on student prior-knowledge and learning experiences in a student-centred learning environment and considers how a student uses learning experiences to construct knowledge (Ertmer & Newby, 2013). For example, studies have shown that comprehension is not just reading information; instead, reading is a constructive activity that involves combining information outside a text, prior-knowledge, and forging a total picture that represents understanding (Ertmer & Newby, 2013; Land, Hannafin, & Oliver, 2012; Spiro et al., 1991). Ertmer and Newby (2013) suggested a principle from the constructivist view for instructional designers: use problem-solving skills that allow learners to go beyond the given information. This constructivist approach is sometimes implemented in problem-based learning. Research has applied the concept of constructivism to advance problem-based learning in the field of educational technology (Dick, 1991).

A learning activity with a chatbot can be seen as a type of problem-based learning because students are required to develop understanding of the chatbot's instruction as a step toward learning how to give useful peer feedback (Barrows & Tamblyn, 1980; Savery, 2006). Reiners et al. (2014) suggested the possibility of designing a chatbot based on constructivism in which the chatbot acts as a mentor, observing and guiding learners to achieve learning objectives; and interfering by initiating a discussion about how to approach an objective rather than dictating how to proceed. The chatbot created for this dissertation, called DD, offers options to guide students developing their reviews of peers' draft outlines. Among these options are concepts about how to give feedback from which a peer can learn.

Ertmer and Newby (2013) proposed a constructivist approach to measuring learning should include how learners engage and perform using a learning tool. Spiro et al. (1991) pointed out that an effective approach to instruction should consider several

highly intertwined issues such as patterns of failures of knowledge-transfer and a theory of learning that addresses known patterns of failures of knowledge-transfer. Therefore, this dissertation study uses concepts of constructivism to investigate student engagement patterns, their perceptions of the chatbot DD, and to propose a methodology about how to use these engagement patterns, including failures, to refine and improve the chatbot's effects.

2.2. The use and evaluation of chatbots

A chatbot is a software program that facilitates a conversation with humans. There are two types of chatbot conversations: non-task-oriented (e.g. chitchat or casual dialogue) and task-oriented (e.g. purchasing a ticket or providing customer support). The origin of a non-task-oriented chatbot can be traced back to ELIZA, a chatbot program developed at MIT, which used natural language to evolve personal conversation with humans (Weizenbaum, 1966). ELIZA used keyword-matching to search a backend template for replies to human input. If an answer could not be found in the backend template, ELIZA used natural language computational methods to reorganize keywords from a user's input and prompt the user to continue the conversation. The purpose of the ELIZA chatbot was to engage a user in a more personal conversation rather than lead the user from a conclusion (Weizenbaum, 1966; Serban et al., 2017).

On the other hand, a task-oriented chatbot is designed to provide instruction or guidance by prompting or nudging a user with information they need to complete an activity (Reiners et al., 2014). This type of chatbot works better in a relatively simple scenario where it tries to keep a user "on track." Tasks explored have included greetings, developing peer reviews, and providing writing assistance about a particular topic (Reiners et al., 2014).

As mentioned in chapter one, there are several ways a chatbot can manage user input (i.e., keyword-matching techniques, machine learning model incorporation, efficiency of conversational exchange) and various indicators of users' satisfaction with a chatbot (i.e., problem-solved and ease of use, chatbot's contribution to the conversation/task, and user interest towards a chatbot). Here I review literature about how chatbots manage novel input from users and how researchers evaluated chatbots in educational settings. My aim was to identify ways a chatbot could play a role in

educational activities and identify kinds of data that should be gathered when designing a chatbot.

2.2.1. Keyword-matching techniques

Sarosa et al. (2020) developed a chatbot for English learners using the Chatfuel package on the Facebook Messenger platform. Its question-answer interaction was based on a keyword-matching technique. The chatbot was able to give quizzes to the students. However, the study did not report user satisfaction or whether learners' English performance improved. As a result, the effect of the chatbot remains unknown.

Ahmad et al. (2020) designed a FAQ chatbot using the keyword-matching techniques and embedded it in a university's website for university marketing purposes. The chatbot could retrieve information, make appointments and registration, and receive feedback from users for improvement. For example, when the chatbot was not able to recognize a phrase from a user, it first triggered a fallback response and prompted the user to give feedback. Then, the user could type "feedback" and add new keywords to the chatbot system. Although the authors mentioned the chatbot was developed using Agile involving qualitative data gathered by interviewing students, the methodology and results are weak as this study did not describe the design process or any results about student success or satisfaction.

Abbasi and Kazi (2014) designed an information retrieval chatbot and studied its effects on memory retention and learning outcomes among 72 sophomore students in a course on object-oriented programming languages. The chatbot's conversational queries were constructed using the Artificial Intelligence Markup Language (AIML) with a knowledge base of terms and keyword-matching techniques. The knowledge base was 5000 question samples collected from the Information Technology Centre, including "what", "why", "how", etc. The students were randomly divided into two groups. One group (n=36) used the Google search engine and the second group (n=36) used the chatbot to search for information to solve programming problems. Retention of information students studied was tested at all three intervals after the problem-solving experience: 5 minutes, 2 hours, or 1 day later. Learning outcomes were measured by a pretest and posttest for which students were given 10 minutes to complete five problem-solving tasks. Students remembered information developed with the chatbot better than

students using the Google search engine. The students in the chatbot group performed better than those in the Google search engine group on the problem-solving tasks. However, this study did not investigate how students interacted with the chatbot. Thus, it is unclear how the chatbot helped students in searching for solutions and what factors improved memory for content when interacting with the chatbot.

Clarizia et al. (2018) evaluated the performance of a chatbot prototype with 187 undergraduate students by investigating system logs and administering a questionnaire in two computer science courses offered on an e-learning platform. The chatbot was based on keyword-matching techniques within its Knowledge Base module. The evaluation included whether the chatbot offered (a) a correct suggestion, (b) a correct suggestion that did not fit with the context, or (c) a wrong suggestion. Initial findings indicated the chatbot was not able to distinguish similar meanings for words, such as C or C++, as the name of a programming language. Moreover, when students introduced into code an argument that had several meanings, the chatbot failed to provide a correct response. Such issues illustrated the need for improving the NLU for future chatbot designers and are also consistent with suggestions by Kerly et al. (2007). On the usability questionnaire, students reported the chatbot was easy to use and user-friendly, however the questionnaire was not provided in this publication.

2.2.2. Machine learning model incorporation

Serban et al. (2017) developed a chatbot using deep learning and reinforcement learning for Amazon Alexa Prize competition. They incorporated several machine learning response models such as knowledge base question answering systems and template-based systems. This method has the advantage of dealing with user input from various domains, such as information retrieval and movie users might want to view. They used A/B testing with real-world users, which allowed them to compare various dialogues and policies by keeping other system factors constant. They found that using deep learning and reinforcement learning techniques could be a way to improve the accuracy of responses from a chatbot if it was provided additional data.

Sinha et al. (2020) used the K-means clustering algorithm as identify groups of data formed on the basis of similar information. They prepared 1000 educational query-based conversations from various websites forming 800 pairs of conversations for the

training dataset and reserving the remaining 200 pairs of conversations for the test dataset. They found that the response accuracy of the chatbot was approximately 60% when K = 4 (F-measure, 57%-60%). When the chatbot failed to answer correctly, the system sent an email to a human responder. However, this study did not identify sources where of the dataset nor evaluate the chatbot with users.

Mekni et al. (2020) developed an information retrieving (FAQ) chatbot using DialogFlow. It was designed to be used in several universities in Minnesota. The purpose of the chatbot was to support academic advising and academic counselling to offer instant support to student, staff, and faculty. Although they sketched the systematic design of their chatbot, the evaluation was not provided. Similarly, Ranavare and Kamath (2020) proposed a FAQ chatbot using DialogFlow for a professional college but the study did not report on the effectiveness of the chatbot.

Goel and Polepeddi (2019) designed Jill Watson, a virtual teaching assistant in an online class on knowledge-based artificial intelligence at Georgia Tech. What inspired them to develop Jill Watson was the challenge of giving timely, individualized, and quality feedback to students. Jill Watson is a FAQ chatbot and can respond to student introductions about themselves. Version 1.0 used IBM Bluemix toolsuite, but it was able to answer only a small percentage of students' questions. Jill Watson 2.0 added semantic processing methods developed in the researchers' laboratory and successfully replied to about 40% of student inputs. Jill Watson 3.0 relied on episodic memory with semantic processing and increased successful replies to 60% of student introductions, and the questions answered by version 3.0 were 91% correct. A larger dataset improved Jill Watson's performance and accuracy. When the researchers revealed Jill Watson's real identity – a chatbot – to students, they were surprisingly positive and felt they were interacting with a human teaching assistant. The design of Jill Watson was domain-specific with using a rather extensive dataset to enhance its performance, which is consistent with Wang and Petrina (2013)'s suggestions.

2.2.3. Efficiency of conversational exchange

Pham et al. (2018) conducted a pilot study with a Google Dialogflow chatbot involving English language learners using a mobile device. The design of the chatbot used predefined dialogue data to communicate with student users. This allowed users to

choose a sentence by clicking a button instead of typing. For example, if the chatbot asks, “How are you?” a list of clickable buttons pop-up, such as “I’m fine,” “I feel terrible,” and “Pretty good.” Pham et al. (2018) indicated that such a design eases interaction with a chatbot. The chatbot was able to make “small talk,” offer learning activities such as quizzes or grammar lessons, explain or provide a hint about those activities, and remind learners to study. The chatbot was deployed on the Google Play store, and 14,000 people participated for two months in the pilot study. Phan et al. analyzed the participants-to-chatbot interaction based on system logs. Results showed there were only 2.3 messages per conversation on average generated by the participants, and some participants left a conversation after two or three messages because the chatbot was not able to handle the participants’ requests. One explanation for this meager finding was that participants were not well-trained, and some of them were new English learners, so their interaction with the chatbot and English understanding were limited. The researchers suggested that detailed instructions should be provided for users, and the design of a chatbot interface should be user-friendly and straightforward. This study offered some design suggestions for researching effects of chatbots, e.g., using predefined context and learning activities. However, this study did not consider users’ backgrounds, such as language proficiency and age, when designing the chatbot. As a result, some users were not able to comprehend materials and interact productively with the chatbot.

2.2.4. Evaluation of a chatbot - problem-solved and ease of use

Kerly et al. (2007) studied 30 students (11 final year undergraduates and 19 graduate students) interacting with a chatbot using the Wizard-of-Oz method to learn C programming. In the Wizard-of-Oz method, a human operator pretends to be a chatbot that interacts with the students. Because the dialogues were expected to be complex, the researchers pointed out that this method can overcome programming errors and would be appropriate for their purposes. The student could view how the system modelled their knowledge, and they could negotiate with the chatbot to change the “system’s belief” (model of the learner) about how well the learner had acquired content. The researchers identified design features important for a chatbot interacting with students in an open learner modelling environment. A chatbot should: (1) being able to connect to a database to update and store data, (2) be able to handle user requests,

such as small talk, to deliver a productive conversation, (3) keep the user on topic, (4) always be accessible, (5) improve the chatbot's natural language understanding unit (NLU) by reviewing/analyzing conversations, and (6) make it easy for users to access the chatbot, e.g., through web integration. In this study, interaction with the chatbot was text-based. It is unclear how integrating other channels, such as audio or graphics, into a chatbot would affect the quality of or increase interaction with students.

2.2.5. Evaluation of a chatbot - chatbot's contribution to the conversation/task

Goda et al. (2014) investigated the effects of a chatbot on a pre-discussion activity before EFL students participated in a human-to-human group discussion. The researchers carried out two case studies involving a total of 130 university students divided into two groups each participating in two successive class periods (case 1: *n*=63; case 2: *n*=67). In both groups, one class period was designated as experimental in which students conversed with a chatbot, and another class period served as a comparison experience in which students listed their thoughts and searched relevant information on the internet. The chatbot was a revision of ELIZA implementing a Socratic dialogue[1] and embedded in Blackboard (an online learning management system). Discussion with the chatbot was intended to serve as a pre-discussion facilitator for students. Case 1 examined the overall effect of a 10- minutes discussion with the chatbot in the first session, followed by a human-to-human online discussion of 30 minutes. The researchers found the experimental group with chatbot made a greater number of conversational contributions than the comparison group in the human-to-human online discussion, but scores reflecting critical thinking skills and satisfaction were the same across groups. Case 2 followed the same procedure as Case 1 to investigate changes in critical thinking using a critical thinking questionnaire pre- and post. Although results from Case 2 showed no overall difference between the experimental and comparison groups, the researchers further analyzed four factors of critical thinking: (1) awareness of critical thinking, (2) inquiring mindset, (3) objectivity, and (4) importance of evidence. While implementing Socratic inquiry might have had a

[1] The Socratic dialogue method is "subject-oriented, rather than subjective. The critical thinker, as an individual and in relations with others, is concerned with exploring issues, rather than with the subjective merging of identities" (Furedy & Furedy, 1985, p. 57).

positive impact on students' critical thinking in terms of awareness of critical thinking and inquiring mindset, the effects of Socratic inquiry on conversations between students and the chatbot were not investigated. Consequently, assigning effects on students' critical thinking to the chatbot or the Socratic inquiry method is confounded. Moreover, some students expressed negative opinions about the chatbot because it sometimes did not speak meaningfully. An example was that sometimes an answer from the chatbot was unrelated to prior conversation. Another issue is the very short span of engagement with the chatbot, just 10 minutes. Nonetheless, the researchers suggested the possibility of increasing student engagement with a chatbot in a pre-discussion activity.

Wang and Petrina (2013) explored how student-produced data can be used to guide the instructional design process of an intelligent language tutoring chatbot called Lucy. The dialogues Lucy generated were based on the match between predefined dialogue templates to user input and were built in AIML on the Pandorabots website. They evaluated Lucy by examining entries in the system's log. In the beginning, Lucy did not respond in a meaningful and accurate manner and was confused by variable learner inputs. To overcome these issues, the researchers refined Lucy by dividing content into small domains using discourse analysis, such as travel, tour guide, and waitress conversation. They also discovered Lucy would repeat the same content with students, which they hypothesized would benefit learners by helping them to better understand sentence structures, as found by Fryer and Carpenter (2006). Also, Lucy provided grammar and spelling checks by gathering input from the learners and added this data into "her" database. In other words, the more learners used Lucy, the more data was gathered, which iteratively enhanced Lucy's performance. However, this required a large volume of data, a potential limitation to developing chatbots with capacities to process the wide diversity of language inputs from users.

Heller et al. (2005) investigated the effects of Freudbot used by 53 students in a 10-minute session during a distance-education psychology course. Freudbot was designed in AIML with keyword-matching techniques to respond to user input and was programmed to chat based on Sigmund Freudian's terminology, concepts, and theories. The researchers' analysis of student-chatbot interactions from the chatlog and responses on five-point Likert items indicated the chat stayed on-task 90% of the time versus drifting off-task or into distractions. Administering a practice quiz in the chatbot was rated the most important feature by the students. The personality of the chatbot was

the second most important design feature. Students mentioned that chatting with Freudbot felt like chatting with Freud and offered several suggestions to improve Freudbot's chat behaviour. Wang and Petrina (2013) concluded developing a successful chatbot depends on a continuously refined process.

Song et al. (2017) integrated a participatory design method to develop a chatbot prototype for two successive graduate-level educational technology online courses (*n*=11). The participatory design method involves all stakeholders (e.g., course instructor, researcher, instructional designer, and software engineer) in designing and developing a system. The students in both courses interacted with the chatbot to reflect on learning materials over 12 weeks of the course. The chatbot was implemented in the first course as a prototype. Then, the researcher used student log data to refine the prototype and deployed it in the second course. A rule-based method selected preplanned sentences as the chatbot's contributions to conversations with the students. The study reported reflection activity with the chatbot was successful in terms of overall participation rate (98.5%) and the reflection task participation rate (80% - 90%). However, how the chatbot affected reflection activity was unclear. How did the chatbot support student reflection? How did students perceive the chatbot? The researchers agreed with Wang and Petrina (2013) that data-based input is needed to enhance a chatbot's performance.

2.2.6. Evaluation of a chatbot - user interest towards a chatbot

Fryer and Carpenter (2006) studied benefits of using chatbots to assist undergraduate students (*n*=211) learning a foreign language. Two chatbots were used in their study: ALICE and Jabberwacky. These chatbots were designed to be entertaining and make conversation. First, students used ALICE for 20 minutes during a class in which they practiced English listening and reading skills. Then, students filled out a questionnaire about their experience with the chatbot although contents of the questionnaire were not presented in their publication. Next, students used Jabberwacky to describe their experience with the ALICE bot. The students felt more motivated and engaged when interacting with both of the chatbots than a human partner. Although these findings seem promising, the study did not investigate effects of the chatbot beyond whether students liked it.

Fryer et al. (2017) investigated students' (*n*=122) task interest with different partners, chatbot or human, in first- and second-year compulsory English as a Foreign Language (EFL) in Japan. The Cleverbot, chatbot software for learning foreign languages and general communication was used in their study. The researchers divided students into two groups to carry out three speaking tasks. Each task lasted two weeks. In the first task of the first week, group A interacted with a chatbot and group B interacted with a human partner. The following week the conditions were reversed and repeated with the same speaking task. One week later, a course interest scale was distributed to the students. In the second task of the first week, a new speaking task was given to the students, and they followed the same research design procedure as for task one but without filling out a course interest scale. Task 3 was also repeated but with a course interest scale. Students' interest when interacting with the chatbot partner decreased from the first to the second task, while those who interacted with a human partner reminded high in all three tasks. The researchers interpreted that practicing speaking with the chatbot might not be an authentic experience for students. Thus, the students saw it as a poor learning experience. Their results further suggested that the character of a task can be critically important when integrating innovative technology, such as a chatbot, into the student learning experience. However, this study did not explain how they designed their chatbot, for example, keywords the chatbot used. I hypothesize chatbot design might influence the students' interacting experience.

Reiners et al. (2014) conducted semi-structured interviews with six educators to gather opinions about using chatbots. The educators identified several reasons they were reluctant to use chatbots in an educational setting. First, building a chatbot is a complex and error-prone process. Secondly, there is a lack of flexibility in chatbot development tools for educators. Current tools require expertise beyond "average" educators or collaborating with a professional in chatbot design and programming. Thirdly, the NLU in chatbots needs to be improved because most users still perceive chatbot language as unnatural during interactions. Lastly, there are limited pedagogical models, and ready-to-use systems with support educators need to integrate a chatbot with an e-learning platform. Reiners et al. suggested pedagogical models, such as behaviourism and constructivism, as guides for incorporating a chatbot into educational settings. For example, in terms of behaviourism, a chatbot should repeat and explain lessons/materials and provide feedback rewards (e.g., praise) for on-task behaviour.

Together, this study identified some obstacles needing to be overcome and introduced pedagogical suggestions to bring a chatbot into authentic educational situations.

In short, several issues were identified in the literature regarding methods for evaluating a chatbot. First, if a conversation with the chatbot lacks meaningfulness, chatbot designers should think about how to re-design the conversation flow and to re-engage students with a chatbot. Second, when interactions with chatbots last a short span by of time (e.g. Goda et al., 2014), it remains unclear whether increasing the exposure time can affect student engagement or perceptions. Third, extensive user data is required to maximize a chatbot's performance. Fourth, the design of a chatbot should be divided into small domains to limit conversational errors. Fifth, how a chatbot affects student learning and how students perceive chatbots is unclear (i.e., Song et al., 2017's study). Last, a task must be selected carefully to increase students' perceptions that engaging with the chatbot is an authentic experience (Fryer et al., 2017).

To sum up, the literature suggests several features for designing an effective chatbot, especially in an educational context. The chatbot should (a) provide conversation meaningfully and communicate with students as humans do; (b) keep students on topic; (c) provide a quiz function to foster learning; (d) integrate graphics to facilitate student engagement; and (e) use dialogue data to continuously refine the NLU and the chatbot itself. Notably, research to date has not often investigated student engagement patterns while interacting with a chatbot. Understanding student-to-chatbot engagement patterns provides valuable instructional insights regarding how students converse with a chatbot and whether conversations support learning needed to meet educational objectives. Giving peer feedback during peer review is a complex writing learning activity, so incorporating a chatbot in this activity might help students to develop more effective feedback. Therefore, this dissertation will try to answer the following three research questions by incorporating technical and other suggestions from the literature to design a chatbot called DD; this dissertation also proposes a methodology for the future chatbot developers and instructional designer(s) to address several the aforementioned needs.

1. How did students interact with the writing chatbot? What were their patterns of engagement?

2. How can writing instructors/instructional designers support peer review activities through innovative chatbot technology?

3. What may improve student-to-chatbot interactions in a discipline-focused writing course?

Chapter 3.

The Design of The Chatbot DD

The chatbot DD was developed using Rasa (version: rasa_core 0.11.12, rasa_core_sdk: 0.11.5, rasa_nlu 0.13.7), an open-source conversational artificial intelligence (AI) framework (Bocklisch, Faulkner, Pawlowski, & Nichol, 2017). Rasa is programmed in the Python programming language. The main reason I chose Rasa is because Rasa meets the needs of non-specialist software developers in the research field (Bocklisch et al., 2017) and its data can be stored on a local machine instead of a shared cloud server, which complies with the province of British Columbia's regulations regarding storage of personal information and data usage. Rasa also provides dialogue management with machine learning to "remember" context by learning from interactions with users, thus increasing training data. Rasa also offers several options for deployment, such as on the cloud or a local server. Moreover, developers have flexibility to customize the design of the chatbot, for example, by integrating customized code into Rasa and database management. However, one major challenge I encountered was immature documentation. Another challenge was the cycle of upgrades to Rasa requiring regular attention to adjusting my code and system architecture.

Rasa has three parts used in chatbot development (Figure 1): (1) Rasa core which consists of Stories (customizable dialogues) and a Domain (the universe in which the bot lives), (2) Rasa core SDK (software development kit) includes customizable actions, such as data storage or database access, and (3) Rasa NLU (natural language unit) which consists of training data for the chatbot to understand utterances within its catalogue.

A web-based server (i.e., chatroom) serves as a bridge between Rasa and users (see Figure 2). A user can either enter text in the textbox and click the "SUBMIT" button (Figure 2) or click a button on the screen (Figure 3) to interact with DD. However, to save the time and to avoid NLU limitations confusing the chatbot (e.g., synonyms or similar phrasing), the students in this study mainly interacted with the DD by clicking buttons (Ahmad et al., 2020; Clarizia et al., 2018; Jain et al., 2018; Luger & Sellen, 2016; Pham et al., 2018). Studies (Clarizia et al., 2018; Reiners et al., 2014) emphasized the

importance of the pedagogical perspective when designing a productive conversational flow to achieve successful interaction experience. Following suggestions by Wang and Petrina (2013) regarding instructional design, a chatbot should be designed for specific domains and utilize users' input to iteratively reshape it into successively better versions. Thus, the chatbot in this dissertation eliminates the chitchat and plagiarism functions and focuses on guiding peer review. The design of the chatbot DD also uses human-like natural language (e.g., praise and causal language) and incorporates several pictures of my dog, which I predicted would increase positive engagement (Jain et al., 2018; Luger & Sellen, 2016). DD's human-like natural language (scripts) were modified by the members at the Ed Psych Lab at SFU to make it more enjoyable.

Another essential feature of chatbot design is the fallback action. A fallback action is triggered when a chatbot cannot process what a user inputs, or a chatbot action fails to be triggered due to student user technical errors.

Jain et al. (2018) noted that designers must explicitly design for conversation failure situations. Figure 2 shows how a user can recover from a conversation error. All the conversation with the chatbot was automatically stored in the backend of Rasa server even if a user restarted the conversation or a failure situation occurred.

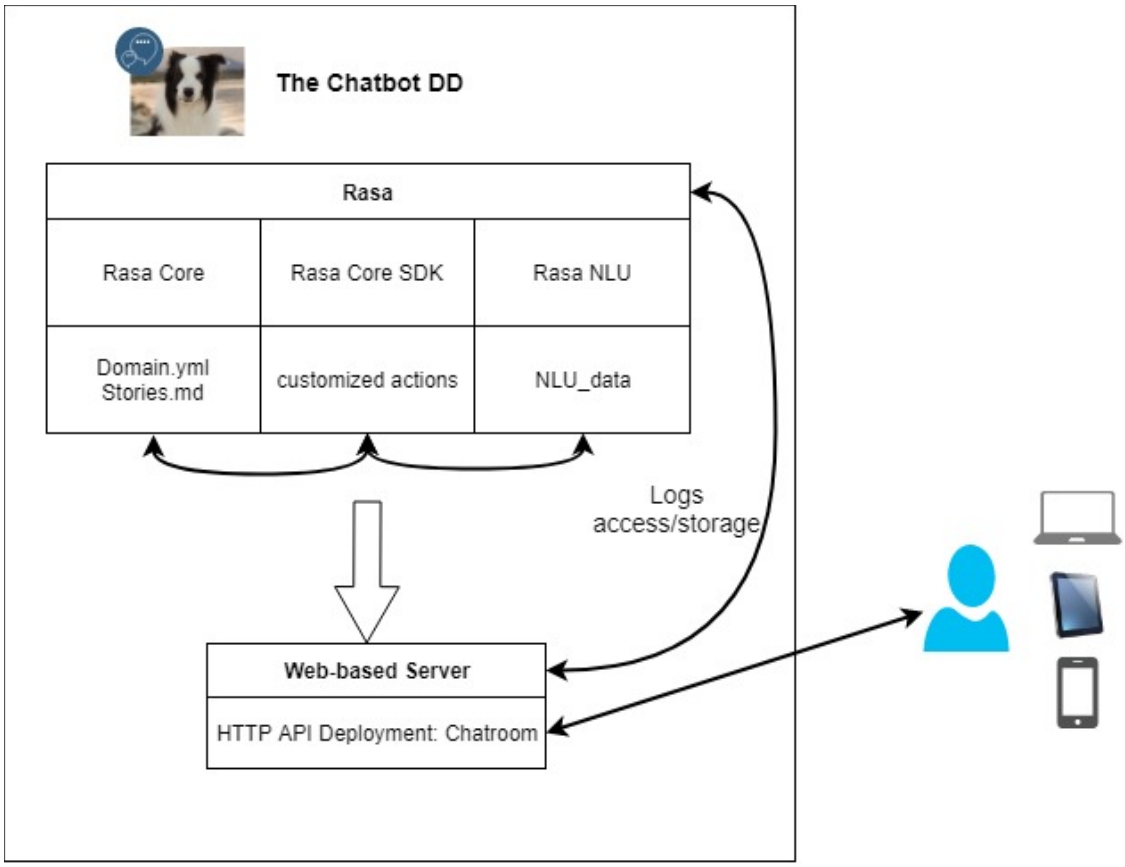

Figure 1. The framework of the chatbot DD.

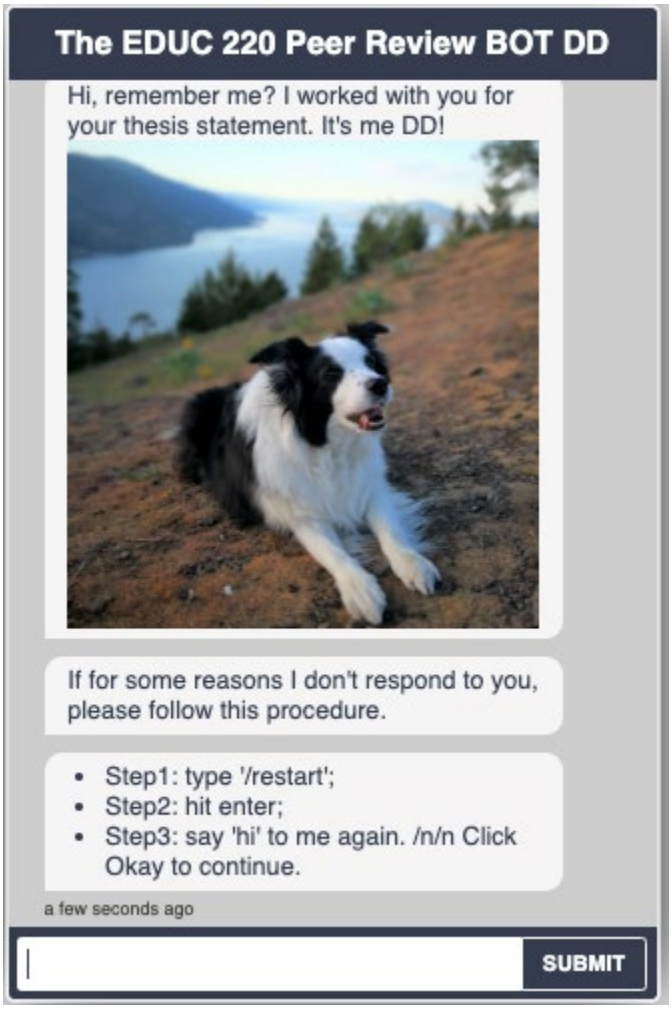

Figure 2. **DD's web-based chatbot interface uses the Chatroom feature.**

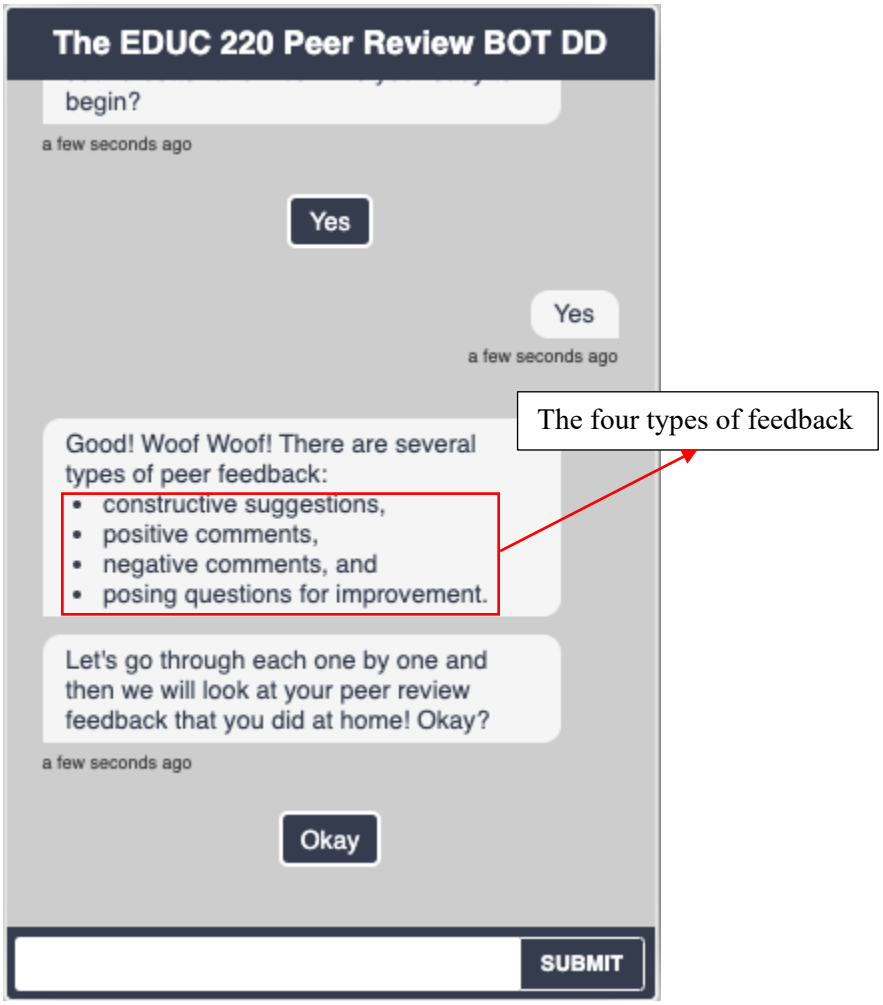

Figure 3. The first session: learning about four types of feedback.

The design of the chatbot DD utilized a participatory design method that involved consultations with the course instructor, a former teaching assistant, and members from an educational research lab at a university in Western Canada who provided feedback to refine the chatbot over a two-month period. Before rolling out the chatbot, a work-study undergraduate student and all members of the lab tested it.

DD first greets students and asks for the student's identification (ID) number. This ID is stored on the backend server for future reference and data analysis. Students participated in two sessions with DD. In the first, they learn how to give effective feedback on a peer's design for an argumentative outline using four types of comments: constructive suggestions, positive comments, negative comments, and posing questions for improvement. In the second session, students review a Dialectical Map (DM) with the chatbot's guidance. The DM is a software tool that organizes and visualizes ideas for an

argumentative essay. A map can serve as the outline for the final argumentative paper in the course (see Nesbit, Niu, & Liu, 2019; Niu, Sharp, & Nesbit, 2015).

The four types of feedback students learned to use as reviewers of a peer's DM were based on Gielen, Peeters, Dochy, Onghena, and Struyven's (2010) research: constructive suggestions, positive comments, negative comments, and posing questions for improvement. Several studies (e.g., Cho & Cho, 2011; Kulkarni et al., 2016; Nelson & Schunn, 2009) investigated the effectiveness of peer feedback to maximize reviewer's critical thinking skills and to improve student writers' mastery of course materials. Effective peer feedback in the form of constructive suggestions and negative feedback should identify specifics about the problem needing attention, provide an explicit correction or an explanation (Cho & Cho, 2011; Gielen et al., 2010; Nelson & Schunn, 2009; Topping et al., 2000). The difference between constructive suggestions and negative comments is the form give the author information about why and how an idea unit or rhetorical structure could be improved. For example, "This sentence needs more elaboration" or "You should change A instead of B because this will make your essay structure better." On the other hand, negative comments point out problem areas regarding essay structure and implicitly tell the author where a misconception is without offering suggestions to rectify it. Positive comments may increase student motivation in learning (Lu & Law, 2012; Topping et al., 2000). When giving positive comments, reviewers should point out why or how an idea makes a good point instead of simply praising. An example is shown in Figure 4. Posing questions for improvement advances student reflection (Gielen et al., 2010; Lan & Lin, 2011; Prins, Sluijsmans, & Kirschner, 2006). Taken together, based on the literature, giving these four types of feedback together may alert reviewers to aspects of quality and informativeness in an essay and help writer's profit from reviews to improve a draft essay.

Figure 4. Learning positive comments through DD the chatbot.

DD's engagement with students about each type of feedback followed a sequential plan. After a student is provided the definition of a type of feedback, DD tests the student's understanding by asking one to two comprehension checking questions (CCQ). Furthermore, to increase student motivation in this learning activity, DD praises a student when s/he correctly answers the CCQs (Song et al., 2017). The CCQs in DD were designed by a former teaching assistant of the course. In real classroom settings, studies showed that CCQs offer teachers the opportunity to check whether students understand content and create opportunities for students to recall content (Chen, Wei, Wu, & Uden, 2009; King, 1994; Redfield & Rousseau, 1981). Including CCQs in DD allows it to mimic real-time and live engagement between teachers and students. An example is provided in Figure 4.

After students have been introduced to and tested on all four types of feedback, they are invited to give feedback on a DM. As shown in Figure 5, students could choose to provide any of the four types of feedback on a thesis statement and accompanying arguments/counterargument with guidance from the chatbot. Students' interactions with DD (e.g., button clicked and texts submitted) were stored on the backend Rasa server.

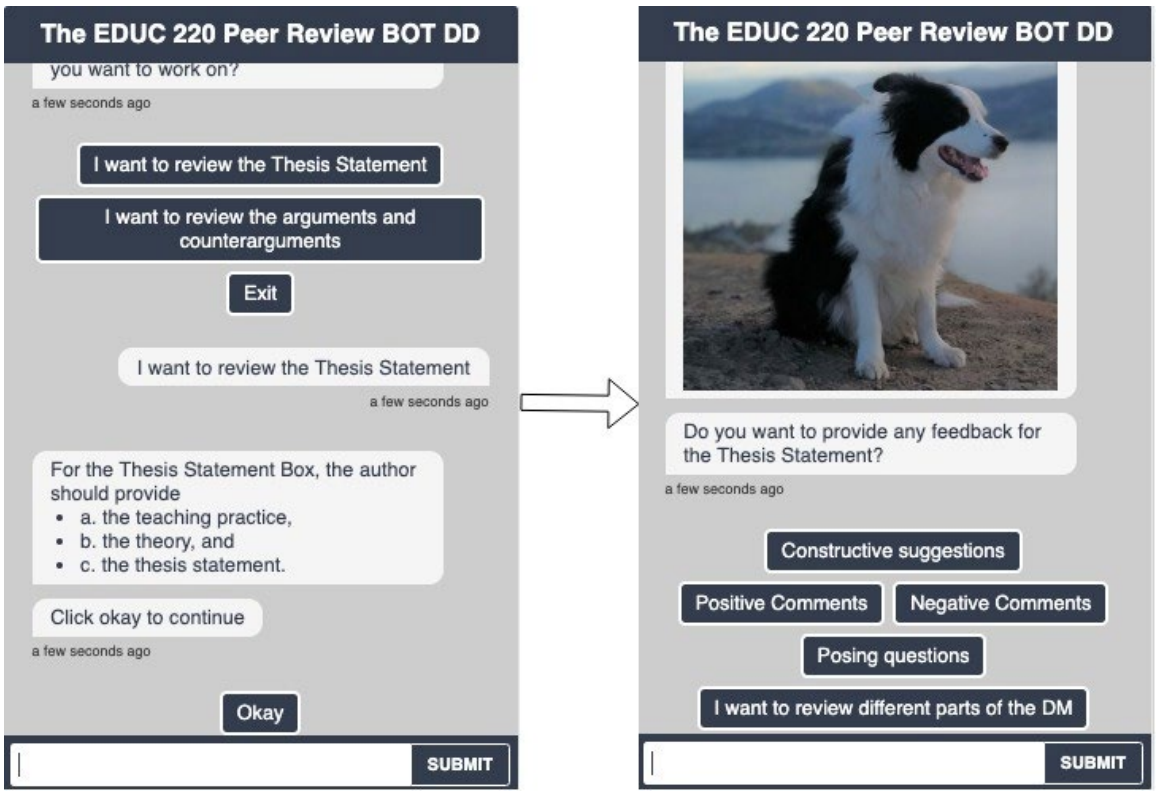

Figure 5. Review thesis statement and arguments and counterarguments with guidance from DD the chatbot

Chapter 4.

Methods

4.1. Participants

Participants were recruited from a tutorial section in a first-year educational psychology course at a university in Western Canada. There were about 200 undergraduate students divided into ten tutorial classes. With their consent, 23 students (n = 23) enrolled in this class participated in this study. Participants academic majors ranged over Arts & Social Science (74%), Education (9%), and Science (17%). In the total sample of 23, 22% of participants were non-native English speakers and 78% were native English speakers.

4.2. Instruments

There are four main instruments in this study: the DM, a review sheet, the chatbot, and a questionnaire. First, the DM was developed by Dr. John Nesbit's research team in the Faculty of Education at SFU. The course instructor decided to use the DM in the course. It is a visualization tool for planning an argumentative essay. In the undergraduate course, students could use the DM to outline their thesis statement, argument, evidence, warrant, counterargument, and conclusion of an essay (Nesbit, Niu, & Liu, 2019; Silverman, 2019).

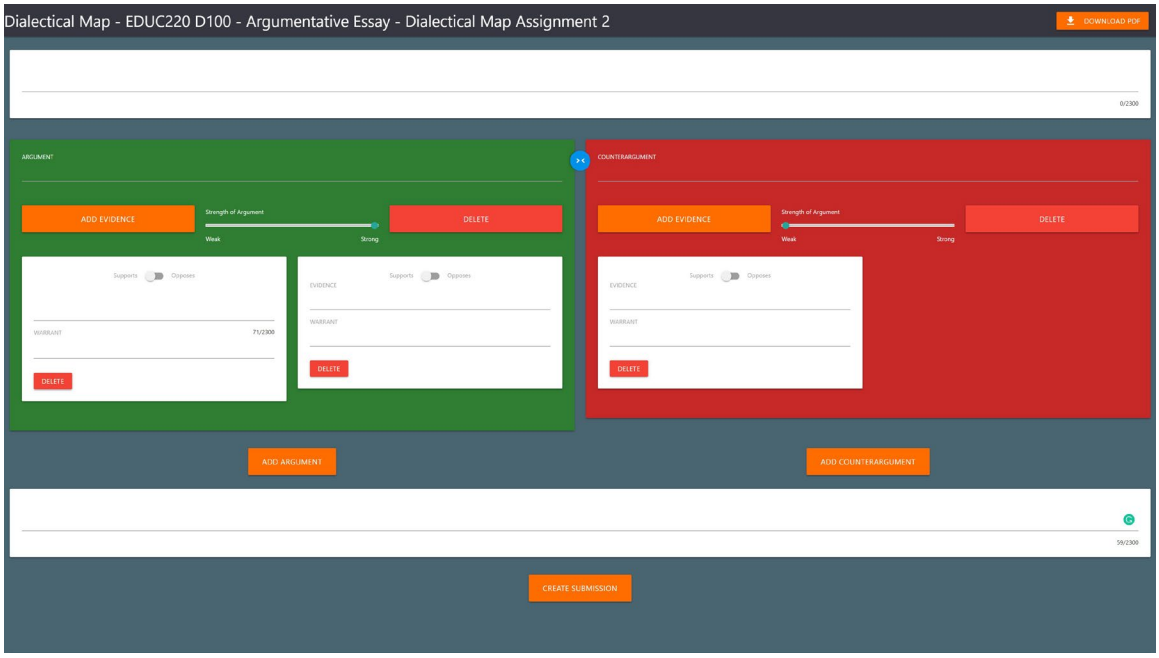

Figure 6. An example of DM.

Collaborating with a colleague who was gathering other data involving this sample of participants, we created five sample argumentative DMs using argumentative essays from students enrolled in the preceding year of this course (Fall 2018). We also created fabricated feedback on a review sheet (i.e., the materials) for each of these DMs. These were base materials given to participants in this study to be improved.

A review sheet with fabricated feedback was distributed to each student. This review sheet contains four types feedback with prompts to guide students in completing the task (see Appendix A).

As described in Chapter 3, DD the chatbot was constructed to assist students in giving more effective feedback in peer review of a DM.

Lastly, a questionnaire (see Appendix B) was adopted and revised from Schunn et al. (2016), Topping et al. (2000), and Torrance, Thomas, and Robinson (1994). The questionnaire measured students' experience with peer review and their engagement with DD.

4.3. Procedure

One of the requirements in the undergraduate course was to develop an argumentative essay. This assignment was assigned towards the end of the semester. The course instructor strongly recommended students plan their essay using the DM, an online visual planning too. A completed DM counted 10% toward the final course grade. The DM tool has been used in several research studies (e.g., Nesbit et al., 2019; Niu, Sharp, & Nesbit, 2015) and in previous offerings of this course.

Figure 6 illustrates the study procedure[2]. In a regular scheduled tutorial class (50 minutes), participants gathered in a computer lab and were asked to improve the fabricated peer feedback while interacting with DD the chatbot. Each participant was randomly assigned one of the five fabricated DMs to and introduced the purposes of the session and the chatbot. Then, the students interacted with the chatbot in two sessions: (a) learning how to give effective peer feedback, and (b) reviewing the fabricated peer feedback on a fabricated DM to improve the feedback.

[2] Ideally and originally, the study was hoping to investigate how effective the bot was. In other words, the students were supposed to complete a peer review homework using a peer review sheet without the chatbot at home before going to the tutorial class. However, only a few students completed the peer review at home. Furthermore, none of the students who consented to participate in this study completed the peer review task at home.

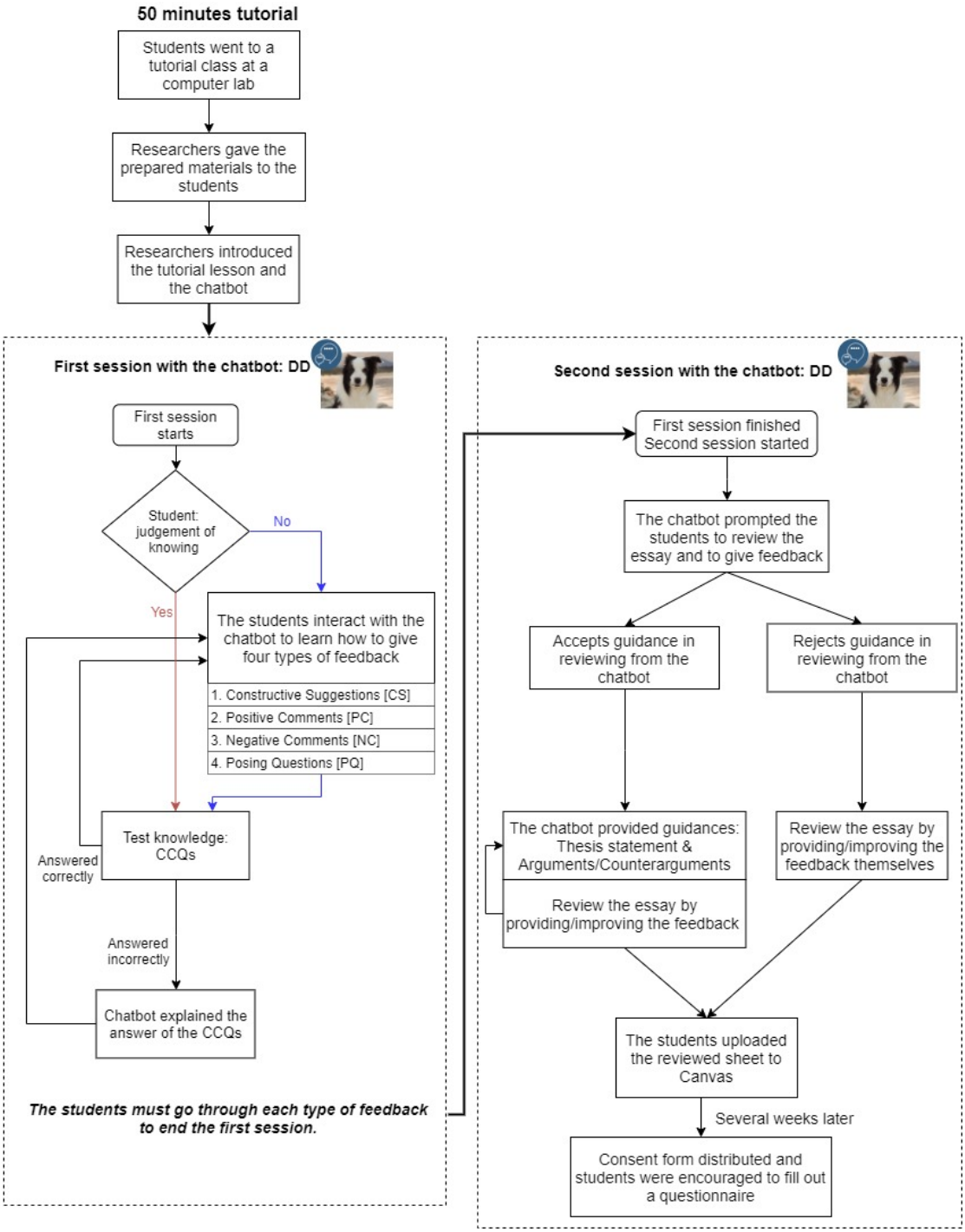

Figure 7. The study procedures.

The chatbot guided participants through self-evaluation and possibly studying four types of peer feedback identified in Figure 3: constructive suggestions, positive comments, negative comments, and posing questions for improvement. First, the chatbot asked the participant whether s/he knew how to provide positive comments.

Then the participant made a self-judgement. If the participant judged they were familiar with a type of feedback, the chatbot tested his/her understanding by posing one to two CCQs. If a participant indicated s/he did not know how to appropriately provide feedback of that type, the chatbot offered instruction followed by a test of the effects of the chatbot's instruction gauged by one to two CCQs. Whenever the participant answered both CCQs correctly, the chatbot offered praise and progressed to the next type of feedback. However, if the participant answered one or both CCQs incorrectly, the chatbot provided the correct answer with an explanation that instructed about that type of feedback, then restarted this cycle for the next type of feedback. This protocol is shown in Figures 7 and 8.

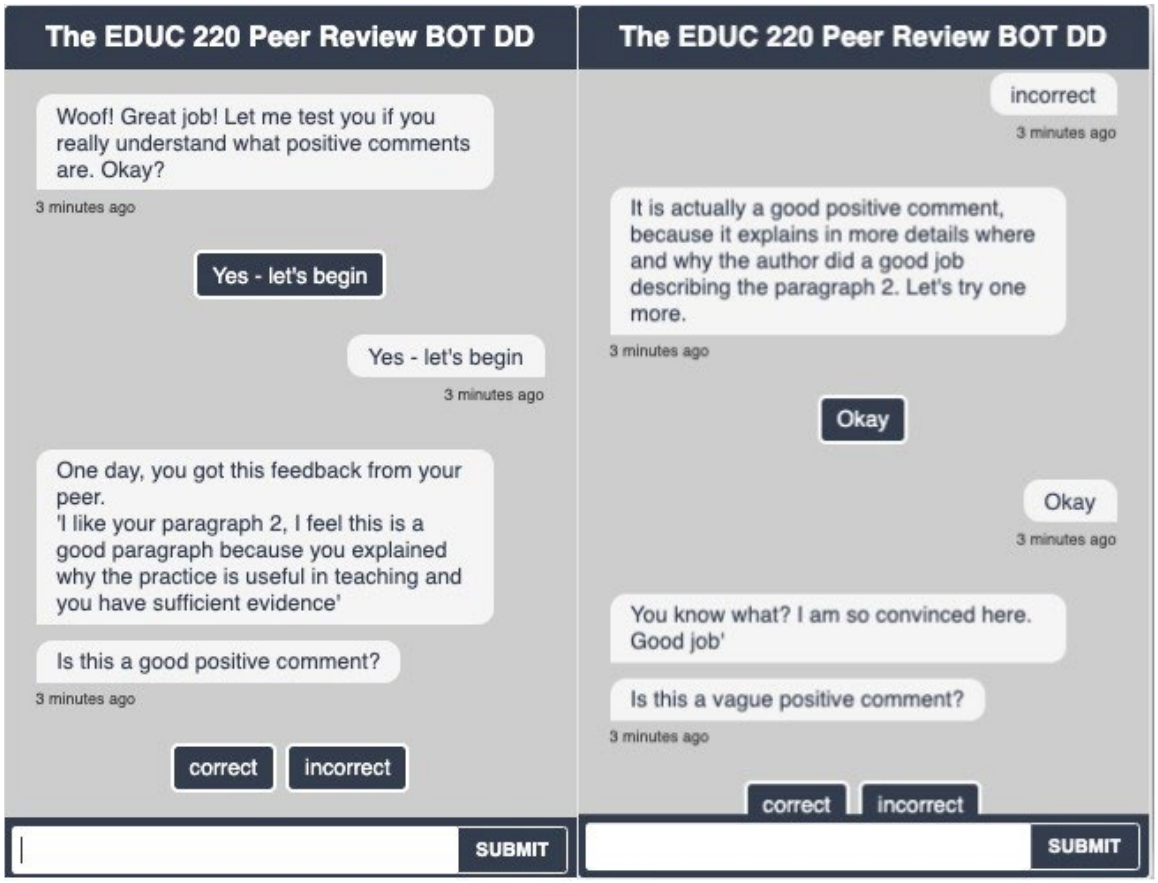

Figure 8. An explanation from the chatbot when the student answered incorrectly.

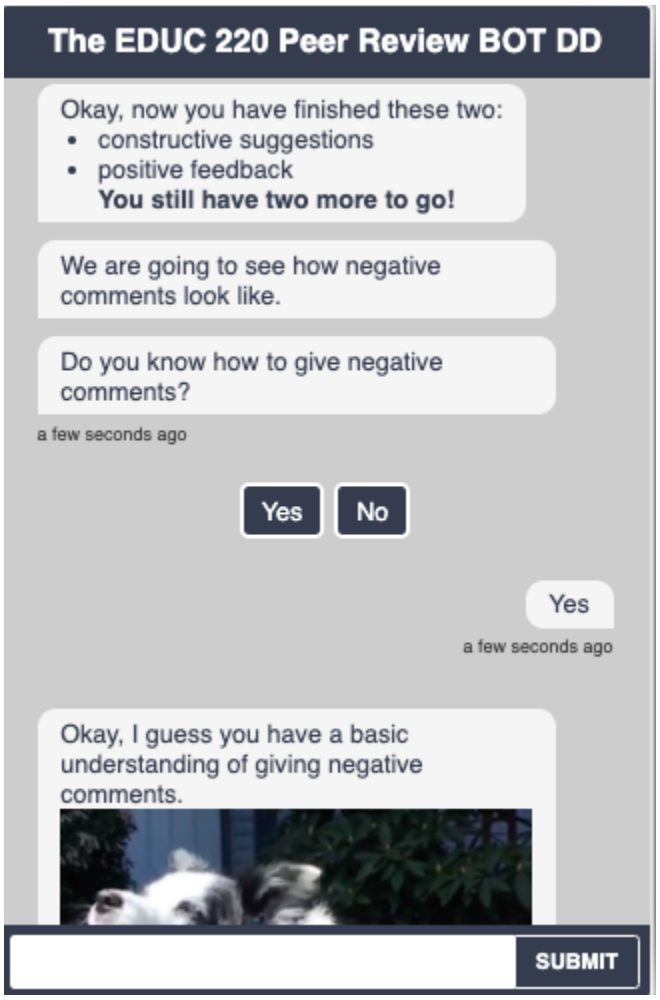

Figure 9. The chatbot will test the student's understanding when s/he judged knowing how to give particular feedback.

Upon completion of the first session, the student then decided whether s/he would like to review the DM with the guidance from the chatbot. If s/he accepted the guidance from the chatbot, the chatbot provided more in-depth guidance about reviewing the DM. As shown in Figure 10, the student indicated that s/he wanted to review constructive suggestions and the chatbot provided prompts to the student. Meanwhile, the student started to review the DM and to improve the fabricated sample feedback on the peer review sheet. These revisions were made using Microsoft Word. After participants finished polishing the feedback sheet, they ended the conversation with the chatbot, automatically causing all the conversation to be saved. Alternatively, if s/he rejected guidance from the chatbot, the chatbot program terminated after all entries in the preceding conversation were automatically saved. The participant was required to upload their improved peer review sheet to Canvas, the course learning management

system. Lastly, towards the end of the semester, the consent form was distributed to the participants, and they had time to fill out a questionnaire regarding their experiences with the chatbot until the semester ended in April. The questionnaire can be found in the Appendix B.

In the second tutorial session, participants analyzed the DM they had been assigned in relation to requirements the instructor had set for the assignment (see Figure 9). As shown in Figure 6, the chatbot was available to provide guidance if the participant asked for it.

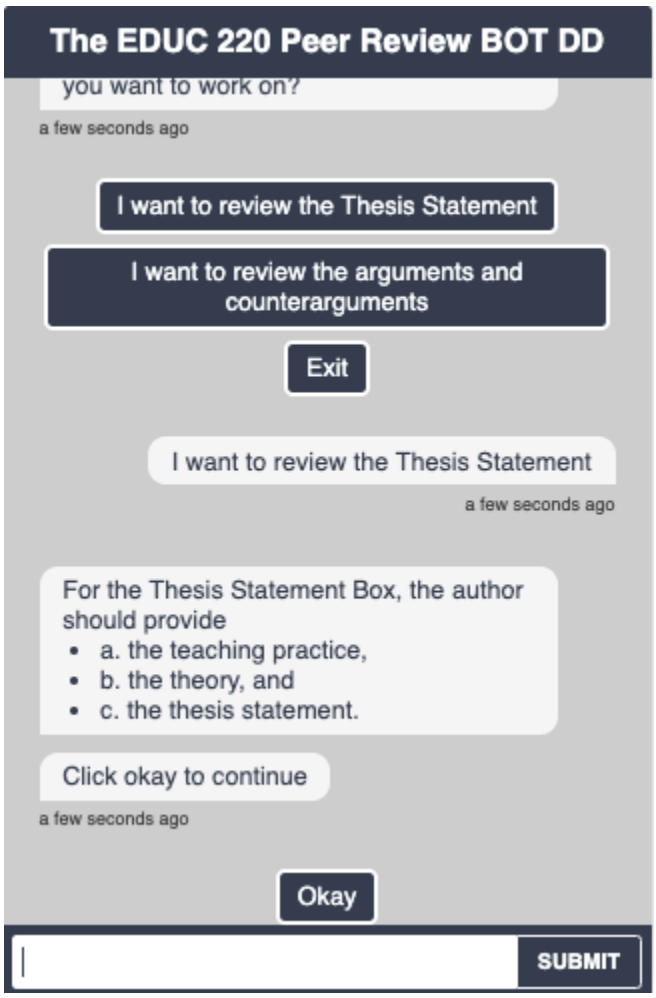

Figure 10. The second session of the chatbot prompts the students to review thesis statement, arguments/counterarguments, or exit the session.

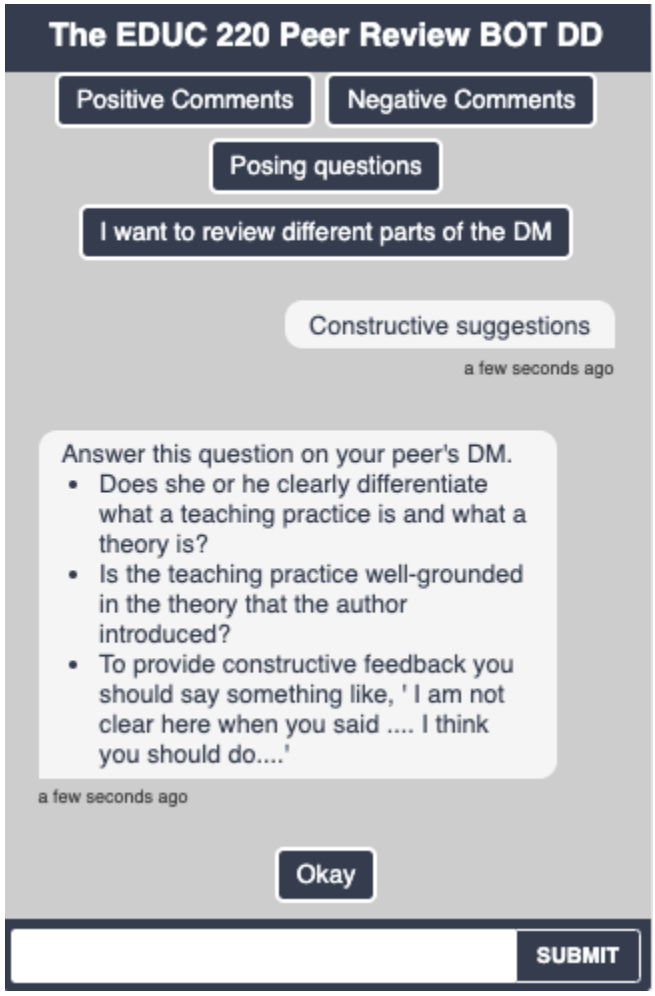

Figure 11. The chatbot prompted the student to review constructive suggestions on thesis statement.

4.4. Data Analysis

Data collected in this study included the participants' recommendations for improving fabricated review on review sheet, a transcript of each participant's chat history with the chatbot, and responses on a questionnaire probing participants' experience with the chatbot. An effective approach to analyze students' engagement patterns is content analysis as this method uncovers and explores student data to generate inferences about student engagement patterns (Chen et al., 2011; Patton, 1990; Weber, 1990; Yang, 2010).

The chat history was stored in json format. I used an online tool (https://jsonformatter.org/) to format the chat history and installed the participant's

choices into an Excel spreadsheet. For instance, self-judgement is coded as 0 (do not know how to give the feedback) and 1 (know how to give the feedback), respectively. Then, an experienced colleague who has previous TA experience in educational psychology and I met to categorize the participants' feedback to explore how students chose to give the feedback with/without guidance from the chatbot in the second session where participants could accept or revise fabricated feedback about the fabricated DM.

This coding process produced a two by two matrix as shown in Table 1. If the participant rejected guidance from the chatbot and correctly revised fabricated feedback, it coded as CR (correct response, rejected guidance). Or, if the participant accepted guidance from the chatbot and correctly made revised fabricated feedback, it was coded as CA (correct response, accepted guidance). After coding, I operationally defined and investigated relationships between interacting with the chatbot and participants' skill to appropriately revise fabricated feedback. Lastly, the students' perceptions of the chatbot were analyzed using their responses to the questionnaire.

Table 1. A summary of choices of guidance from the chatbot and its possible interpretations

	Revision correct (C)	**Revision incorrect (I)**
(R) Rejects Guidance from the chatbot (Exit)	CR - Prior session effective	IR – Participant misjudges learning from the prior session
(A) Accepts Guidance from the chatbot	CA - Participant judges the need help, get it and succeed at revising insufficient peer feedback	IA - Chatbot instruction in the prior session and in the second activity not effective

CR means the student made the revision correct and rejected guidance from the chatbot. IR means the student made the revision incorrect and rejected guidance from the chatbot. CA means the student made the revision correct and accepted guidance from the chatbot. IA means the student made the revision incorrect and accepted guidance from the chatbot.

Chapter 5.

Results

5.1. Participant self-judgement and CCQs on learning four types of feedback

A first step in planning analyses of data gathered according to the protocol outlined in the preceding chapter is to characterize inferences regarding students' knowledge about providing good feedback. These "possible paths" across states of participants' engagement step from a judgement of knowledge about a particular form of feedback and responses to CCQs that gauge participants' understanding of the four types of feedback. For each of three types of feedback – positive comments, constructive suggestions, and negative comments – eight possible paths, labeled A to H, are catalogued in Table 2. Participant judgement and the correctness of a CCQ about posing questions for improvement generated four possible paths, I to L, as shown in Table 3.

As shown in Table 2, paths B, D, and G indicate the chatbot conversation could be effective in developing knowledge about providing three types of feedback. In paths B and D, the students misjudged their prior-knowledge. Paths C and F match, but are not totally valid evidence, the students had a lucky guess on one of the CCQs. These paths are not very illuminating about effects of the chatbot. Paths A and E might signal the chatbot is ineffective for students who needed instruction. Path H indicates the student may not need chatbot because prior knowledge was sufficient.

Table 2. Possible paths relating participant judgement and CCQs correctness on constructive, positive, and negative feedback after interacting with the chatbot

Path	The paths for constructive suggestions, positive comments, negative comments			Possible Interpretation(s) about chatbot instruction (CI)
	Self-Judgement	CCQs 1 answered correctly	CCQs 2 answered correctly	
A	No	No	No	CI[1] is ineffective
B	No	No	Yes	CI could be effective
C	No	Yes	No	Cannot measure whether CI is effective; lucky guess on the first CCQ
D	No	Yes	Yes	CI could be effective, and the student could know how to give the feedback but judged did not know, the student might have prior knowledge
E	Yes	No	No	CI is ineffective; self-judgement is wrong; other content or other CI contingencies might interfere
F	Yes	Yes	No	Cannot measure whether CI is effective; lucky guess on the first CCQ.
G	Yes	No	Yes	Self-judgement is somehow wrong, but CI could be effective
H	Yes	Yes	Yes	The participant may not need CI; participant has prior-knowledge

[1] CI symbolizes the chatbot instruction

Table 3 maps possible paths for posing helpful questions as a form of feedback. Path J shows the chatbot instruction appears to be effective for students who judge prior knowledge insufficient. Paths I and K suggest chatbot does not benefit students. Path L indicates the student already has knowledge about providing helpful questions to improve an essay and may not need to chat with the chatbot.

Table 3. A summary of student judgement and CCQ correctness on posing questions for improvement by interacting with the chatbot

	Path for pose a question		Possible Interpretation(s)
Path	Self-Judgement	CCQ answered correctly	
I	No	No	CI[1] is ineffective
J	No	Yes	CI could be effective; the participant could know how to pose questions but judged did not know
K	Yes	No	Self-judgement is wrong; prior-knowledge misconception; CI could be ineffective
L	Yes	Yes	The participant may not need CI; participant has prior-knowledge

[1] CI symbolizes the chatbot instruction

Tables 4 to 7 show counts of participant whose data match the possible paths just described. Each path comes with possible interpretations on constructive suggestions, positive comments, negative comments, and posing question for improvement. In Table 4, data indicate the chatbot instruction could be effective for 8 out of 23 participants when learning constructive suggestions (paths B+D+G). However, the chatbot instruction could be ineffective for 4 participants when learning constructive suggestions (paths A+E). In addition, 9 participants might have made a lucky guess on the CCQs (paths C+F).

Table 4. Summary of interaction paths of constructive suggestions and possible interpretations

Path	Constructive Suggestions	Possible Interpretation(s)
A	3	CI[1] is ineffective
B	3	CI could be effective
C	7	Cannot measure whether CI is effective; lucky guess on the first CCQ
D	4	CI could be effective, and the participant could know how to give the feedback but judged did not know, the participant might have prior knowledge
E	1	CI is ineffective; self-judgement is wrong; other content or other CI contingencies might interfere
F	2	Cannot measure whether CI is effective; lucky guess on the first CCQ.
G	1	Self-judgement is somehow wrong, but CI could be effective
H	2	The participant may not need CI; participant already prior knowledge

[1] CI symbolizes the chatbot instruction

In Table 5, 10 out of 23 participants may not need chatbot instruction on feedback of the form positive comments (path H). Also, 6 out of 23 participants paths reflect an effective chatbot conversation (paths B, D). Nevertheless, 7 out of 23 participants following paths C or F appear to have made a lucky guess when learning about positive comments by interacting with the chatbot.

Table 5. A summary of interaction paths of positive comments and possible interpretations

Possible Path	Positive Comments	Possible Interpretation(s)
A	0	CI[1] is ineffective
B	1	CI could be effective
C	6	Cannot measure whether CI is effective; lucky guess on the first CCQ
D	5	CI could be effective, and the participant could know how to give the feedback but judged did not know, the participant might have prior knowledge
E	0	CI is ineffective; self-judgement is wrong; other content or other CI contingencies might interfere
F	1	Cannot measure whether CI is effective; lucky guess on the first CCQ.
G	0	Self-judgement is somehow wrong, but CI could be effective
H	10	The participant may not need CI; participant already prior knowledge

[1] CI symbolizes the chatbot instruction

Table 6 describes participants when learning about feedback in the form of negative comments by interacting with the chatbot and its interpretations. The majority of participants (*n*=17) showed the effectiveness of chatbot instruction (paths B, D, G). One participant appears to have made a lucky guess (path C), and one other participant showed the chatbot instruction might be ineffective (path E).

Table 6. A summary of interaction paths of negative comments and possible interpretations

Path	Negative Comments	Possible Interpretation(s)
A	0	CI[1] is ineffective
B	7	CI could be effective
C	1	Cannot measure whether CI is effective; lucky guess on the first CCQ
D	7	CI could be effective, and the participant could know how to give the feedback but judged did not know, the participant might have prior knowledge
E	1	CI is ineffective; self-judgement is wrong; other content or other CI contingencies might interfere
F	0	Cannot measure whether CI is effective; lucky guess on the first CCQ.
G	3	Self-judgement is somehow wrong, but CI could be effective
H	4	The participant may not need CI; participant already prior knowledge

[1] CI symbolizes the chatbot instruction

In Table 7, more than half of the participants (*n*=16) reflected the effectiveness of the chatbot when learning how to pose questions for improvement (path J). Only two participants showed the chatbot instruction might be ineffective (path I). Finally, 5 out of 23 participants may not need guidance from the chatbot about posing questions for improvement because they may already have prior knowledge (path L).

Table 7. A summary of interaction paths of posing questions for improvement and possible interpretations

Path	Pose A Question	Possible Interpretation(s)
I	2	CI[1] could be ineffective
J	16	CI could be effective; the participant could know how to pose questions but judged did not know
K	0	Self-judgement is wrong; prior-knowledge misconception; CI could be ineffective
L	5	The participant may not need CI; participant has prior knowledge

[1] CI symbolizes the chatbot instruction

5.2. The second session: Participant revision choices with/without guidance from the chatbot

5.2.1. Students choosing to revise thesis statements

Table 8 describes profiles of participants' engagement with the chatbot in terms of their success in applying each of the types of feedback to revise the thesis statement presented in a fabricated DM. As shown, many participants rejected guidance from the chatbot and proposed incorrect revisions (IR). Particularly in this category, 18 of 23 participants judged they knew how to pose questions for improvement but did not succeed in realizing that kind of feedback. Also,13 of 23 participants incorrectly revised negative comments on a thesis statement without guidance from the chatbot. Furthermore, for the revision of constructive suggestions on thesis statement, 11 out of 23 participants did not correctly revise the fabricated feedback after rejecting guidance from the chatbot.

Not many participants judged they needed guidance, took it and succeeded in revising the fabricated feedback on about thesis statement. For instance, only 2 out of 23 participants chatted with the chatbot about revising negative comments and posing

questions for improvement. However, over one-third of the participants (*n*=9) correctly revised positive comments with guidance from the chatbot.

In terms of the participants who benefited from instruction the chatbot provided in the prior session, more than one-third of participants (*n*=9) correctly revised positive comments and rejected guidance from the chatbot. There are few students with this profile when it comes to revising construction suggestions, negative comments, and posing questions for improvement.

Few participants fell into a group who did not benefit from chatbot instruction in session 1 and unsuccessfully revised the fabricated feedback about a thesis statement. None of the participants who incorrectly revised positive comments had accepted guidance from the chatbot. However, a small number of participants incorrectly revised constructive suggestions, negative comments, and posing questions for improvement who also accepted guidance from the chatbot.

Table 8. Review thesis statement on four types of feedback guided from the chatbot and its interpretations

Thesis statement guidance from the chatbot					
Categories	**Constructive Suggestions**	**Positive Comments**	**Negative Comments**	**Posing Questions**	**Possible Interpretation(s)**
CR – Prior sessions effective	3	9	5	1	CI[1] could be effective due to the former session effects; CI is not needed because the participant already has prior-knowledge
IR – Participant misjudges learning from the prior session	11	5	13	18	Cannot measure whether CI[1] is effective; CI could be ineffective in the prior session
CA – Participant judges the need help, get it and succeed at revising insufficient peer feedback	5	9	2	2	CI[1] is effective
IA – Chatbot instruction in the prior session and the second activity not effective	4	0	3	2	CI[1] could be ineffective
Total	23	23	23	23	

CR means the participant made the revision correct and rejected guidance from the chatbot. IR means the participant made the revision incorrect and rejected guidance from the chatbot. CA means the participant made the revision correct and accepted guidance from the chatbot. IA means the participant made the revision incorrect and accepted guidance from the chatbot.

[1] CI symbolizes the chatbot instruction

5.2.2. Students choosing to revise arguments and counterarguments

Table 9 shows a summary of the numbers of participants regarding the revisions on arguments and counterarguments in relation to the four types of feedback. Compared to revising the feedback on a thesis statement previously, a smaller number of participants when revising arguments and counterarguments are in the IR category. Nevertheless, there are more than one-third of the participants when revising positive and negative comments on arguments and counterarguments in the IR category.

Generally, the CR category points out that over one-third of the participants correctly revised the four types of feedback and rejected guidance from the chatbot on arguments and counterarguments. This result could be an indicator that the prior CI was effective, or the participants already have prior-knowledge.

The category CA reveals a few participants chose to accept guidance from the chatbot and thus correctly revised on positive comments (*n*=4) and negative comments (*n*=6) on arguments and counterarguments. There are also more participants in this category when revising constructive suggestions and posing questions for improvement on arguments and counterarguments comparing to a thesis statement.

Lastly, in the IA category when revising the feedback on arguments and counterarguments, a small number of participants incorrectly revised the feedback and accepted guidance from the chatbot.

Table 9. Review arguments and counterarguments on four types of feedback guided from the chatbot and its interpretations

Arguments and counterarguments guidance from the bot					
Possible Categories	**Constructive Suggestions**	**Positive Comments**	**Negative Comments**	**Posing Questions**	**Possible Interpretation(s)**
CR – Prior sessions effective	9	6	8	10	CI[1] could be effective due to the former session effects; CI is not needed because the participants already has prior-knowledge
IR – Participant misjudges learning from the prior session	4	11	8	3	Cannot measure whether CI[1] is effective; CI could be ineffective in the prior session
CA – Participant judges the need help, get it and succeed at revising insufficient peer feedback	8	4	6	9	CI[1] is effective
IA – Chatbot instruction in the prior session and the second activity not effective	2	2	1	1	CI[1] could be ineffective
Total	23	23	23	23	

CR means the participant made the revision correct and rejected guidance from the chatbot. IR means the participant made the revision incorrect and rejected guidance from the chatbot. CA means the participant made the revision correct and accepted guidance from the chatbot. IA means the participant made the revision incorrect and accepted guidance from the chatbot.

[1] CI symbolizes the chatbot instruction

5.3. Examinations of Students' Review Ability and their Engagement Patterns

Section 5.1 and 5.2 show the effects of the first and second sessions of the chatbot, respectively. In this section, a combination of section 5.1 and 5.2 is depicted in Figures 11 through 14 about participants' engagement patterns when by interacting with the chatbot about four types of feedback revisions. The basic question is, if the chatbot instruction was effective in the first session (paths B, D, or G), did these participants correctly revise with/without the chatbot guidance in the second session (categories: CR or CA)?

5.3.1. Constructive Suggestions

Figure 11 illustrates the participants' engagement patterns when revising constructive suggestions. There is one participant on path B who correctly revised using constructive suggestions after accepting guidance from the chatbot about revising both thesis statement and arguments and counterarguments (CA). There is one participant on path D who correctly revised the fabricated feedback without guidance from the chatbot about both thesis statement and arguments and counterarguments (CR). One participant on path D did not correctly revise the fabricated feedback with chatbot guidance on thesis statement (IA), but correctly revised the fabricated feedback without chatbot guidance on arguments and counterarguments (CR).

Two participants on path D failed to revise the fabricated feedback after rejecting the chatbot's guidance on both thesis statement and arguments and counterarguments (IR). Moreover, a participant on path G successfully revised the fabricated feedback and s/he accepted the guidance from the chatbot on thesis statement (CA), but unsuccessfully revised the fabricated feedback by rejecting the chatbot's guidance on arguments and counterarguments (IR).

Although two participants judged they understood constructive suggestions and correctly answered both CCQs (path H), they both did not correctly revise the fabricated feedback on thesis statement (IR and IA, respectively) while they correctly revised feedback on arguments and counterarguments (CR and CA, respectively).

On the other hand, it is hard to measure how effective chatbot instruction is on paths C and F because participants could have made a lucky guess on the CCQs. The patterns represented by paths C and F show there was only one participant on each path who correctly revised the fabricated feedback on both thesis statement (CA) and arguments and counterarguments (CR). Lastly, it is unexpected that two participants on path A and one participant on path E correctly revised arguments and counterarguments with/without chatbot guidance because both paths A and E could mean the chatbot instruction was ineffective in the first session.

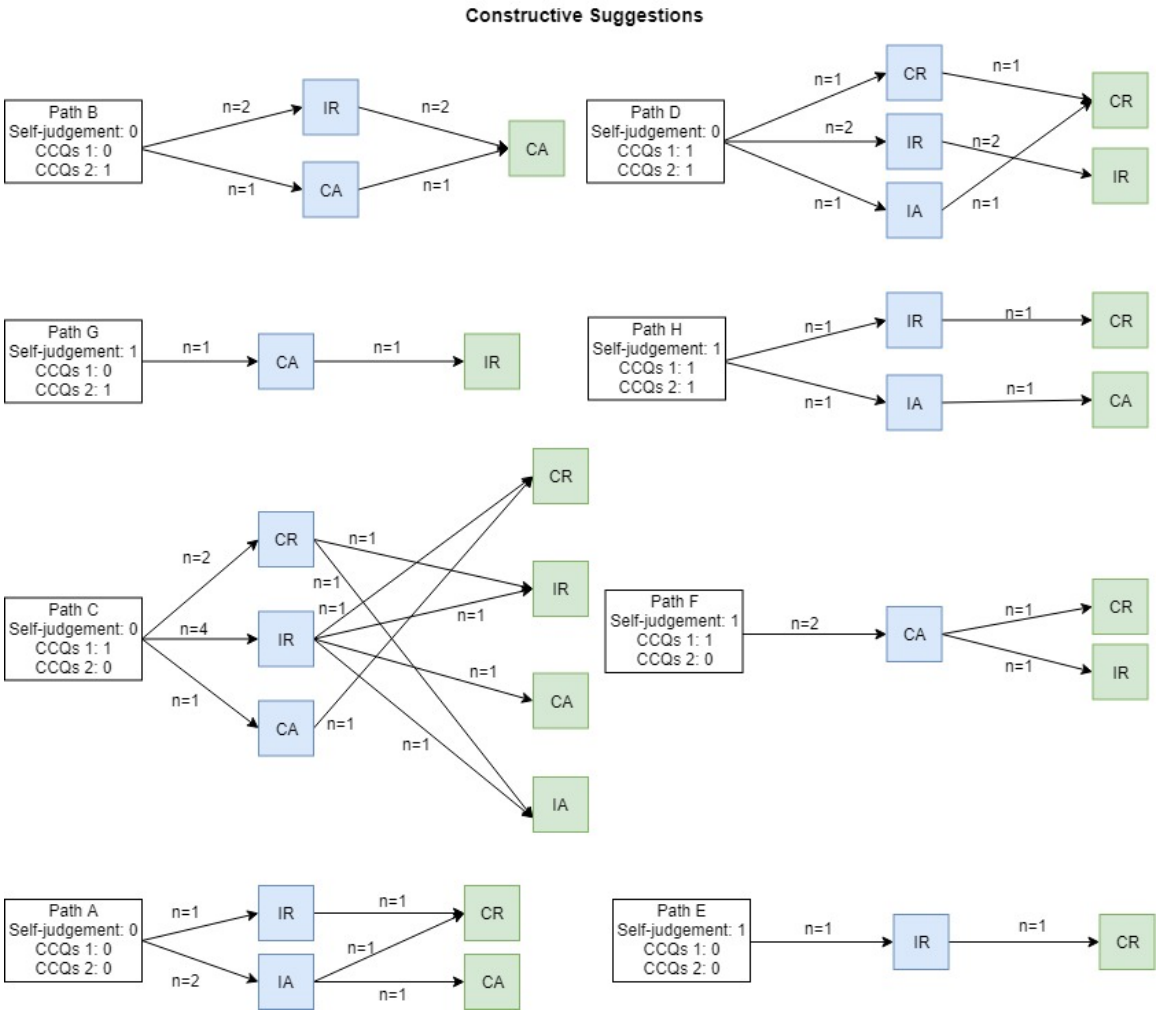

Figure 12. Students' engagement patterns by interacting with the chatbot on constructive suggestions revision

The paths A to H represents the paths of the first chatbot session from section 5.1. The blue square represents the patterns of second chatbot guidance of revision on thesis statement from section 5.2. The green square represents the patterns of second chatbot guidance of revision on arguments/counterarguments from section 5.2. CR means the participant made the revision correct and rejected guidance from the chatbot. IR means the participant made the revision incorrect and rejected guidance from the chatbot. CA means the participant made the revision correct and accepted guidance from the chatbot. IA means the participant made the revision incorrect and accepted guidance from the chatbot.

5.3.2. Positive Comments

In terms of the engagement patterns on positive comments (see Figure 12), the participant on path B incorrectly revised the fabricated feedback without chatbot's guidance on thesis statement (IR) but correctly revised it with chatbot's guidance on arguments and counterarguments (CA).

As shown in Table 2, the chatbot instruction could be effective in the first session (path D). Five participants successfully revised the fabricated feedback, but they rejected the chatbot's guidance on the thesis statement (CR). However, one participant unsuccessfully revised the fabricated feedback without chatbot guidance on arguments and counterarguments (IR). Furthermore, it seems that participants on path H did not know how to revise positive comments. There were only two participants correctly revised the fabricated feedback on both thesis statement and arguments and counterarguments (i.e., blue square CR to green square CR and blue square CA to green square CR). When the participants had a lucky guess (path C and F), only two participants correctly revised on both thesis statement and arguments and counterarguments (blue square CA to green CR).

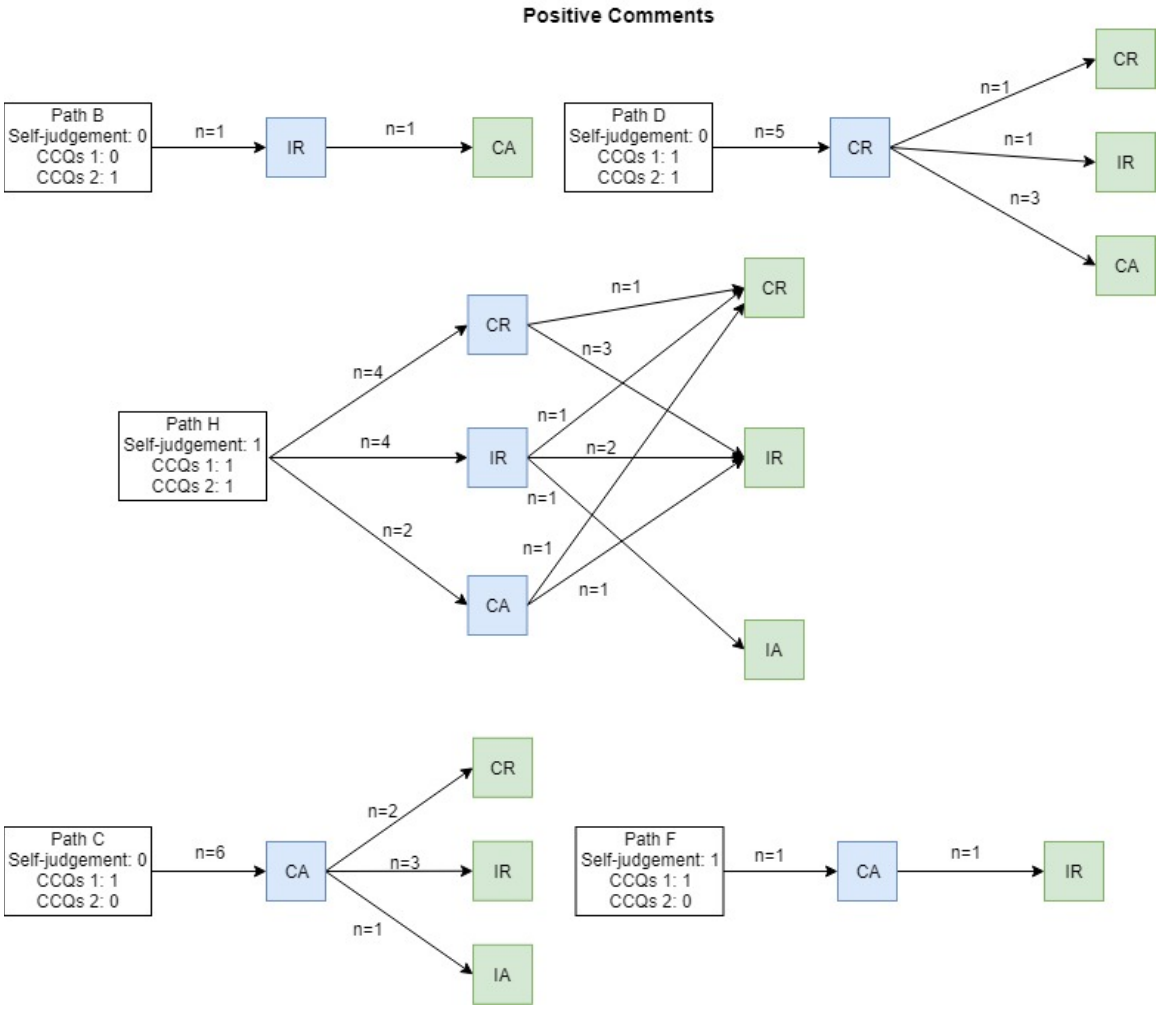

Figure 13. Students' engagement patterns by interacting with the chatbot on positive comments revision

The paths A to H represents the paths of the first chatbot session from section 5.1. The blue square represents the patterns of second chatbot guidance of revision on thesis statement from section 5.2. The green square represents the patterns of second chatbot guidance of revision on arguments/counterarguments from section 5.2. CR means the participant made the revision correct and rejected guidance from the chatbot. IR means the participant made the revision incorrect and rejected guidance from the chatbot. CA means the participant made the revision correct and accepted guidance from the chatbot. IA means the participant made the revision incorrect and accepted guidance from the chatbot.

5.3.3. Negative comments

Figure 13 shows the participants' engagement patterns for negative comments. On path B, when the participants (*n*=4) chose not to accept guidance from the chatbot, they correctly revised the fabricated feedback on the thesis statement (CR). Three of them (*n*=3) also correctly revised on the feedback without the chatbot's guidance on arguments and counterarguments (CR). Although one participant correctly revised the

fabricated feedback on thesis statement with chatbot's guidance (CA), the participant did not correctly revise the feedback on arguments and counterarguments (IR). One participant incorrectly revised the fabricated feedback on thesis statement with chatbot's guidance (IA), but correctly revised it with chatbot guidance on arguments and counterarguments (CA).

Path D shows none of the participants correctly revised the fabricated feedback on both thesis statement and arguments and counterarguments. Although the chatbot instruction might demonstrate its effectiveness in the prior session (path G), three participants incorrectly revised the fabricated feedback, and they rejected the chatbot's guidance on both thesis statement and arguments and counterarguments (IR). Also, when the participants (*n*=4) judged they knew negative comments and correctly answered both CCQs (path H), none of them correctly revised the negative comments on thesis statements and arguments and counterarguments. One participant on path E only correctly revised the fabricated feedback with chatbot guidance on arguments and counterarguments. Nevertheless, when a participant had a lucky guess in the prior session (path F), s/he successfully revised the fabricated feedback on both thesis statement (CA) and arguments and counterarguments (CR).

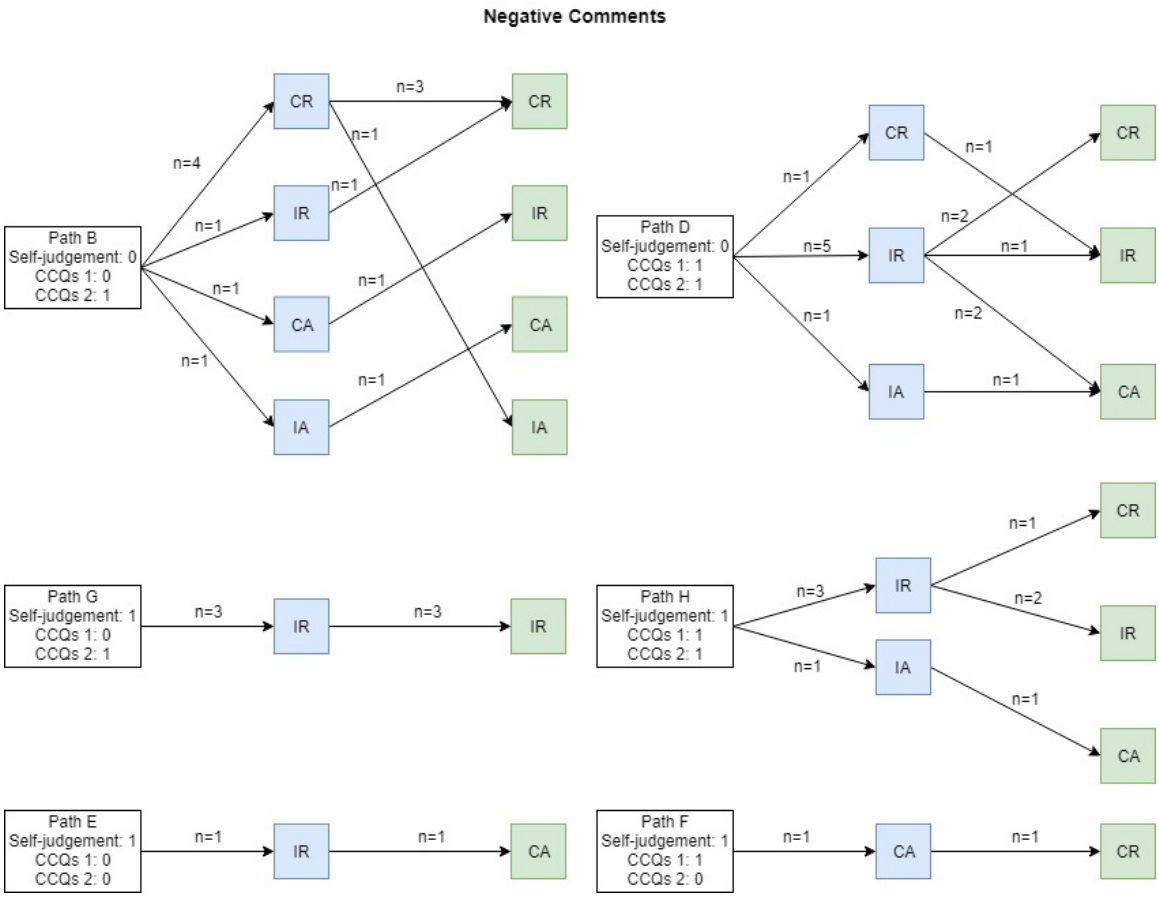

Figure 14. Students' engagement patterns by interacting with the chatbot on negative comments revision

The paths A to H represents the paths of the first chatbot session from section 5.1. The blue square represents the patterns of second chatbot guidance of revision on thesis statement from section 5.2. The green square represents the patterns of second chatbot guidance of revision on arguments/counterarguments from section 5.2. CR means the participant made the revision correct and rejected guidance from the chatbot. IR means the participant made the revision incorrect and rejected guidance from the chatbot. CA means the participant made the revision correct and accepted guidance from the chatbot. IA means the participant made the revision incorrect and accepted guidance from the chatbot.

5.3.4. Posing Questions for Improvement

In Figure 14, path J shows that the majority of participants (*n*=15) incorrectly revised the fabricated feedback when they rejected the chatbot's guidance on thesis statement (IR). However, only eight participants correctly revised the fabricated feedback without chatbot's guidance (CR). Five participants correctly revised the fabricated feedback when they accepted the chatbot's guidance (CA) on arguments and counterarguments. There was only one participant who correctly revised the fabricated feedback on both thesis statement (CR) and arguments and counterarguments (CA).

On path L, only one participant successfully revised the fabricated feedback and accepted the chatbot's guidance on both thesis statement (CA) and arguments and counterarguments (CA). Finally, it is unusual that when the prior session of the chatbot instruction was ineffective (path I), a participant correctly revised the fabricated feedback and s/he accepted the chatbot's guidance on thesis statement (CA). This participant correctly revised the fabricated feedback but rejected the chatbot's guidance on arguments and counterarguments (CR).

All in all, these engagement patterns show various possible explanations and suggestions for future instructional design. I will situate these findings and methodology in the context of discussions, further recommending instructional design methodologies with the integration of a chatbot into educational settings.

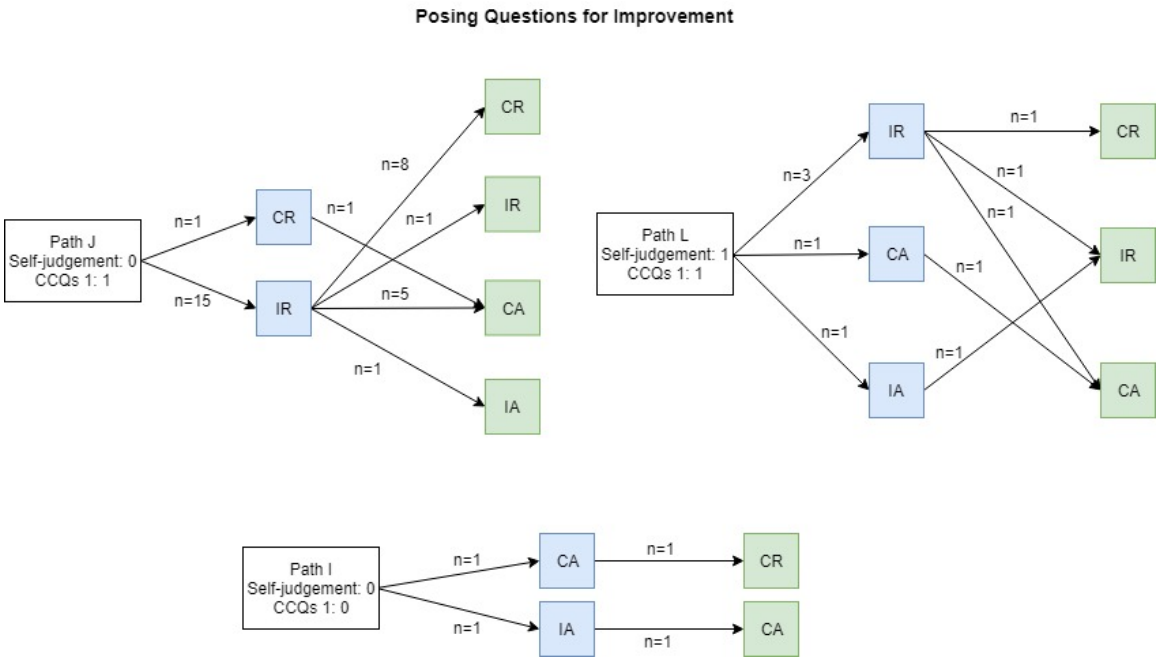

Figure 15. Students' engagement patterns by interacting with the chatbot on posing questions for improvement revision

The paths I to L represents the paths of the first chatbot session from section 5.1. The blue square represents the patterns of second chatbot guidance of revision on thesis statement from section 5.2. The green square represents the patterns of second chatbot guidance of revision on arguments/counterarguments from section 5.2. CR means the participant made the revision correct and rejected guidance from the chatbot. IR means the participant made the revision incorrect and rejected guidance from the chatbot. CA means the participant made the revision correct and accepted guidance from the chatbot. IA means the participant made the revision incorrect and accepted guidance from the chatbot.

5.4. Student Perceptions about the Chatbot

What were participants' experiences? How did they feel about interacting with the chatbot? To investigate participant perceptions toward the chatbot, participants completed a questionnaire.

Table 10 summarizes the results. Twenty participants indicated the chatbot guided them to identify new issues in terms of improving feedback to a peer about the DM. Nineteen participants agreed the chatbot improved their understanding about how to give feedback. Nineteen participants mentioned that interacting with the chatbot DD was enjoyable. Nineteen participants felt that, after interacting with the chatbot DD, they became a better reviewer. There were fourteen participants who found the chatbot DD made the peer review instruction clear. However, a few participants expressed negative opinions about the chatbot DD. They indicated they would rather seek support from professors or teaching assistants.

Table 10. Student perception toward peer review guided by the chatbot

	Experienced Peer Review	Identified New Issues	Improved Feedback	Enjoyable	Became Better Reviewer	Resolved Confusions From The Instructions
Yes	30.43% (7)	86.96% (20)	82.6% (19)	82.6% (19)	82.6% (19)	60.9% (14)
No	47.83% (11)	13.04% (3)	17.4% (4)	17.4% (4)	17.4% (4)	17.4% (4)
N/A	21.74% (5)	0% (0)	0.0% (0)	0.0% (0)	0.0% (0)	21.7% (5)
Total	23	23	23	23	23	23

Table 11 shows participants' perception of the effectiveness of writing by engaging with the chatbot. There were 56.5 % of participants (*n*=13) who expressed that the chatbot guided them to structure the review feedback and 13% of participants (*n*=3) pointed out they improved their writing after engaging with the chatbot. Yet, 4.3% of the participants (*n*=1) reported the chatbot did not help on his/her writing. There were 26.1% of participants (*n*=6) who did not answer this question.

Table 11. The effectiveness of writing by engaging with the chatbot DD

How Did (or Did not) Review Change Because Engaged With The Chatbot	
Structure Review Feedback	56.5% (13)
Writing Improvement	13.0% (3)
Did not Help	4.3% (1)
N/A	26.1% (6)
Total	23

Here are some of the participants' comments:

… interactive, clear Instructions, and an easy-to-follow format

… can answer by clicking a button instead of typing the answer

… the dog pictures made it enjoyable, not stressful, and user-friendly

… promotes students to be critical and reflective in their opinions and ideas

…liked the questions DD asked because they made me think more.

Appendix C contains all student comments about DD. Overall, the participants had positive experiences with the chatbot. Generally, they reported the chatbot guided them in developing review feedback and improved the clarity and detail of their feedback and the presence of animal figures (e.g., my dog) made the experience enjoyable, and DD was easy to interact with. However, one primary suggestion for improvement was to shorten response time. Some participants felt DD's response time was slow as they preferred a faster pace of conversation. Some participants also would prefer DD to ask open-ended questions and provide more examples of feedback they use to learn.

Chapter 6.

Discussion and Recommendation

What has been missing in research about student interaction with chatbot is an instructional design for implementing a writing chatbot in education (Reiners et al., 2014; Wang & Petrina, 2013). To fill this gap in the research, this study proposed a methodology for studying how students engage with a chatbot during a peer review activity, one essential step in a writing process (Cho & Schunn, 2007; Cho & Cho, 2011; Min, 2005). Specifically, this study sought to create, present and test a new methodology for understanding of student-to-chatbot interactions and student engagement patterns when working with a chatbot. The specific context for this work was a peer review activity in an undergraduate course where students were required to use a software tool, the Dialectical Map, to plan an argumentative essay. Another focus of this dissertation was how students perceived the chatbot and what might be improved about student-to-chatbot interactions.

Q1: How did students interact with the writing chatbot? What were their patterns of engagement?

Possible paths were proposed to describe how participants can interact with a chatbot when presented a choice to learn about four forms of feedback: constructive suggestions, positive comments, negative comments and posing questions for improvement. It is difficult to draw conclusions about the effectiveness of a chatbot's instruction just by examining results like those presented in Tables 2 and 3. For instance, a participant who learned about a form of feedback, successfully completed the CCQs, suggests the effectiveness of chatbot instruction (paths B, D, G, or H). However, findings presented in section 5.3 suggest such a participant may not be able to give effective feedback in a genuine task, such as reviewing a DM in the second session of this study. An example appears in Figure 13 where none of the participants from path D correctly revised the fabricated feedback on both thesis statement and argument and counterargument. These findings suggest participants often misjudged their understanding about how to provide feedback. If such participants are given options to engage with a chatbot, they may choose ineffectively.

Schwartz (1994) described judgements like those learners make about optional engagements with a chatbot like DD as “the process of making a prospective judgment at the time of retrieval” (p. 364). Consider the participants’ engagement patterns regarding constructive suggestions in Figure 11 as an example. Two participants who traveled path F indicated that they had known how to provide constructive suggestions; however, answers to the CCQs showed only partial understanding, and one participant did not correctly revise constructive suggestions on arguments and counterarguments. Another example is illustrated by the engagement patterns for feedback of the form negative comments in Figure 13. One participant on path E indicated s/he knew how to address negative comments but failed to correctly answer both CCQs. This participant also did not correctly revise negative comments on thesis statement.

Results like these represent an illusion of knowing. Pashler et al. (2007) characterized the illusion of knowing as “thinking you know something when you don’t” (p. 24). Research has commonly found students typically struggle to make accurate judgements about their learning (Glenberg, Wilkinson, & Epstein, 1982; Pashler et al., 2007). Thus, when designing a chatbot for educational purposes, the instructional designer or researcher must consider how to capture students’ decision making in testing the chatbot because students’ decisions often may be inaccurate.

Q2: How can writing instructors/instructional designers support peer review activity through innovative chatbot technology?

Findings just presented may raise another question: Is the current approach to the instructional design of a chatbot using true/false CCQs adequate? Is just asking a student to judge whether they know how to develop a particular form of feedback a sufficient indication? The purpose of implementing CCQs in the chatbot was to check whether students understood content and to check prior learning (Chen et al., 2009; King, 1994). Roediger III and Karpicke (2006) point out that testing, such as by CCQ, can positively affect student learning, and this testing effect may encourage continuous engagement with the learning materials. Similarly, O'Dowd (2018) found that students saw quizzes used to check knowledge as a formative tool for learning versus just “tests,” and the more quizzes they attempted, the better their online engagement as measured by task completion. Furthermore, from the constructivist view of instructional design, “the focus of the assessment should be on what has been constructed by the learner as a

result of the learning situation" (Dick, 1991, p. 42). Consequently, based on constructivism, students construct knowledge by practicing recall when posed CCQs and through learning experiences in the form of discussions with a chatbot. Based on the findings in this study, we should ensure students can complete a given task successfully beyond just asking them to judge whether they can do that. Thus, CCQs may need to be actual tasks students complete rather than asking for students' perceptions, as a chatbot can do.

Another suggestion from this study's findings may help overcome students' illusion of knowing. It is to detect students' decision paths and provide support when intervention by a chatbot shows benefits. For example, in the future, when a chatbot detects a student travelling path F in the first session, when learning how to develop feedback, the chatbot could redirect the student to study background materials to increase odds the student understands the concept. An instructor/instructional designer can also look at the data describing student-chatbot interactions and plan to offer just-in-time assistance. By analyzing students' engagement behaviour, a chatbot can adapt from the data to match students' knowledge or skills to facilitate learning (Kerly et al., 2007; Wang & Petrina, 2013).

Other findings from investigating students' choices to review and engagement patterns in the second chatbot session indicated that some participants choose to review (e.g., thesis statement and arguments and counterarguments in the DM) with or without guidance from the chatbot. These results generated four categories: CR (correct revision without chatbot guidance), IR (incorrect revision without chatbot guidance), CA (correct revision with chatbot guidance), and IA (incorrect revision with chatbot guidance), respectively.

As can be seen from both Tables 8 and 9, some students did not accept guidance from the chatbot and they incorrectly revised the feedback on DM (category: IR). Such a result is further evidence of students' illusion of knowing. This finding implies possible benefit for an interactive chatbot that allows students to re-engage with the technology might be helpful for learning. A student choosing "No need for guidance from chatbot" might be invited to share revised feedback with the chatbot interface. If the chatbot detects the student's revised feedback is incorrect, the chatbot would provide

corrective guidance. An instructor reviewing the log of such an exchange could also scaffold students.

A few students were accepting of guidance from the chatbot about reviewing the DM although these students did not revise feedback correctly (category: IA). In this case, one may conclude that the chatbot instruction is ineffective. Even with guidance from the chatbot, these students were not able to appropriately revise feedback. This re-confirms need to include CCQs in the context of chatbot design.

Q3: What may improve student-to-chatbot interactions in a discipline-focused writing course?

To sum up, the study investigates the students' engagement patterns by examining the relationships between the two chatbot sessions. From the results in chapter 5.3, we can see that when the chatbot was effective in the first session (e.g., path B), some participants did not successfully revise the fabricated feedback in the second session. To successfully revise the fabricated feedback, the participants must give four types of effective feedback on both thesis statement and arguments and counterarguments. However, not all of the participants gave four types of feedback on both thesis statement and arguments and counterarguments. This could be because the instruction was not explicit; the instruction on the review sheet and the chatbot did not tell the participants that they must give feedback on both thesis statement and arguments and counterarguments. Future instructional designers/researchers are recommended to provide explicit instruction on a task (e.g., peer review).

To prevent student-to-chatbot conversation leads to failure, this study structures the flow of conversational design of the chatbot. The participants mainly interacted with the chatbot by clicking buttons rather than freely making conversations. Studies (e.g., Kerly et al., 2007; Reiners et al., 2014) suggested that better NLU and feedback of an intelligent tutorial system may facilitate better learning outcomes of students. Jain et al. (2018) further suggested that a chatbot must proactively ask good questions to reduce the search space of NLU and engage users in a meaningful conversation. Good questions asking by a chatbot may refer to training data that has been constructed in a chatbot database. Consequently, future chatbot designers will need to improve the NLU mechanism on the classification of synonyms or a similar phrasing on a topic.

Furthermore, the design of chatbot needs progressive refinement that depends on users' data. Prior studies have illustrated ways to use users' input to understand student-to-chatbot interaction and to improve the chatbot (Pereira & Díaz, 2018; Picciano, 2012; Wang & Petrina, 2013). This study added new approaches to this work.

Regarding the students' perceptions of the chatbot, the majority in this study reported positive attitudes. As described in the literature (e.g., Jain et al., 2018; Luger & Sellen, 2016), interactions with a chatbot should be simple and enjoyable because such features of chatbot design may enhance student engagement and learning. Some students pointed out that interacting with the chatbot was enjoyable, and perceived the chatbot improved their learning experience. Students did not feel they were interacting with just a robot or machine because the chatbot replied in "human-like" natural language along with enjoyable pictures of my dog that increased motivation (Fryer & Carpenter, 2006; Jain et al., 2018). In the self-report data, some students also mentioned the chatbot helped them to structure review feedback and improve writing although these effects remain to be validated. Future research should examine both short-term and long-term effects of chatbot on student learning.

In summary, future chatbot designers are recommended to follow several recommendations to improve methodology when designing a chatbot for an educational setting:

(1) Redesign CCQs to be more informing about students' knowledge beyond what can be revealed by true/false questions.

(2) Since some students in this study misjudged learning, a chatbot should be able to test students' understanding extensively and regularly (O'Dowd, 2018; Roediger III & Karpicke, 2006).

(3) Improve the NLU to manage a wider range of conversational forms (Reiners et al., 2014; Wang & Petrina, 2013).

(4) Enhance ability of a chatbot to detect errors students make and lend support to correct errors.

(5) Ensure easy accessibility to lend assistance with a task, for example, as Kerly et al. (2007) suggested, embed the chatbot window or sidebar within a webpage or application window.

(6) Progressive refinement from user input that help to improve chatbot performance and enhance student learning and engagement (Pereira & Díaz, 2018; Wang & Petrina, 2013).

The methodology explored in this dissertation was illustrated in relation to an essay's thesis statement, arguments and counterarguments as represented in a graphical display of argument structure, the DM. I recommend generalizing to other forms of essays, such as lab reports or expository essays. Future research also should explore generalizations of the methodology developed here to investigate data-driven decision making that can further enhance effects of a chatbot by utilizing learning analytics (Picciano, 2012).

6.1. Limitations

This dissertation took an initial step toward developing a methodology to increase understanding about and bring an innovative technology (i.e., chatbot) into educational settings. There are several limitations to this dissertation. First, the small sample size available limits understanding about student engagement patterns in relation to the particular chatbot used in this research. As well, the small training set of thesis statements (about 200) limited the NLU's accuracy.

Possible NLU restrictions or limitations may have misclassified similar sentences that confuse a chatbot (Clarizia et al., 2018).

An example is:

Intent question: "What is a thesis statement?"

Intent thesis_statement: "My thesis statement is positive reinforcement is beneficial to student learning"

Intent thesis_statement_clarification: "Is my thesis statement positive reinforcement may beneficial learning a well-structured thesis statement?".

Such an example may confuse the chatbot as the *intent thesis_statement_clarification* might be identified as the *Intent question* or *Intent thesis_statement* because such sentences are similar and would cause errors when talking to a chatbot. Suggested by one of the developers from the Rasa team, the more training data provided, the fewer errors would occur. We provided about 200 thesis statements as the training data. However, after we trained the NLU model, the chatbot still confused and failed to identify the intent.

This dissertation did not examine reasons for judgements or decisions students made. This information could be beneficial for understanding whether and how student judgement contributes to performance (Bol & Hacker, 2012). Future research might include methods to explore students' reasons for their judgements and choices.

6.2. Future Work

Here I provide some future directions for chatbot research. Some students misjudged their understanding about how to give feedback in this study. Future research could investigate whether increasing the intensity of CCQs, frequency of testing, or reengaging student in learning from a chatbot would repair this problem. Second, future work is recommended on improving a chatbot's accuracy to identify and classify intents. This work would aim to improv the NLU by (a) assembling a large-scale repository of student writing to mark issues and solutions in essays (Kerly et al. 2007; Lin & Chang, 2020; Wang & Petrina, 2013), (b) applying semantic processing based on conceptual representations of knowledge (Goel & Polepeddi, 2019), (c) utilizing a self-repair model with native/non-native speaker chat data (Höhn, 2017), and (d) incorporating a linguistic discourse tree (Galitsky & Ilvovsky, 2017). Third, research may explore how students learn with guidance provided by a chatbot about other sectors of an essay (e.g., body paragraphs) and other types of essay, e.g., expository or narrative.

References

Abbasi, S., & Kazi, H. (2014). Measuring effectiveness of learning chatbot systems on Student's learning outcome and memory retention. *Asian Journal of Applied Science and Engineering*, *3*(2), 251-260.

Ackerman, J. M. (1993). The promise of writing to learn. *Written communication*, *10*(3), 334-370. https://doi.org/10.1177/0741088393010003002

Ahmad, N. A., Baharum, Z., Hamid, M. H. C., & Zainal, A. (2020). UNISEL Bot: Designing Simple Chatbot System for University FAQS. *International Journal of Innovative Technology and Exploring Engineering*, *9(2)*, 4689-4693. DOI: 10.35940/ijitee.B9067.129219

Barrows, H. S., & Tamblyn, R. M. (1980). *Problem-based learning: An approach to medical education*. Springer Publishing Company.

Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here?. *Frontiers in psychology*, *3*, 229. https://doi.org/10.3389/fpsyg.2012.00229

Chang, D. (2020). *What's within a Thesis Statement? Exploring Features of Argumentative Thesis Statements* (Unpublished doctoral dissertation). Simon Fraser University, Burnaby, Canada.

Chen, Y. L., Liu, E. Z. F., Shih, R. C., Wu, C. T., & Yuan, S. M. (2011). Use of peer feedback to enhance elementary students' writing through blogging. *British Journal of Educational Technology*, *42*(1), E1-E4. https://doi.org/10.1111/j.1467-8535.2010.01139.x

Chen, N. S., Wei, C. W., Wu, K. T., & Uden, L. (2009). Effects of high level prompts and peer assessment on online learners' reflection levels. *Computers & Education*, *52*(2), 283-291. https://doi.org/10.1016/j.compedu.2008.08.007

Ching, Y. H., & Hsu, Y. C. (2013). Peer feedback to facilitate project-based learning in an online environment. *The International Review of Research in Open and Distributed Learning*, *14*(5). https://doi.org/10.19173/irrodl.v14i5.1524

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, *39*(5), 629-643.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, *48*(3), 409-426. https://doi.org/10.1016/j.compedu.2005.02.004

Clarizia, F., Colace, F., Lombardi, M., Pascale, F., & Santaniello, D. (2018, October). Chatbot: An education support system for student. In *International Symposium on Cyberspace Safety and Security* (pp. 291-302). Springer, Cham. https://doi.org/10.1007/978-3-030-01689-0_23

Dick, W. (1991). An Instructional Designer's View of Constructivism. *Educational Technology*, *31*(5), 41-44.

Ertmer, P. A., & Newby, T. J. (2013). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance improvement quarterly*, *26*(2), 43-71

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovations in Education and training International*, *32*(2), 175-187. https://doi.org/10.1080/1355800950320212

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior*, *75*, 461-468. https://doi.org/10.1016/j.chb.2017.05.045

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, *10*(3), 8-14.

Furedy, Christine, & Furedy, John J. (1985). Critical thinking: Toward research and dialogue. *New Directions for Teaching and Learning*, *23*, 51-69

Galitsky, B., & Ilvovsky, D. (2017, April). Chatbot with a discourse structure-driven dialogue management. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 87-90). Association for Computational Linguistics.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and instruction*, *20*(4), 304-315. https://doi.org/10.1016/j.learninstruc.2009.08.007

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, *10*(6), 597-602. https://doi.org/10.3758/BF03202442

Goda, Y., Yamada, M., Matsukawa, H., Hata, K., & Yasunami, S. (2014). Conversation with a chatbot before an online EFL group discussion and the effects on critical thinking. *The Journal of Information and Systems in Education*, *13*(1), 1-7. https://doi.org/10.12937/ejsise.13.1

Goel, A. K., & Polepeddi, L.. (2019). Jill Watson. In C. Dede, J. Richards, & B. Saxberg (Eds.) *Learning Engineering for Online Education* (pp. 120-143). New York: Routledge. https://doi.org/10.4324/9781351186193-7

Heller, B., Proctor, M., Mah, D., Jewell, L., & Cheung, B. (2005, June). Freudbot: An investigation of chatbot technology in distance education. In *EdMedia+ Innovate Learning* (pp. 3913-3918). Association for the Advancement of Computing in Education (AACE).

Höhn, S. (2017, August). A data-driven model of explanations for a chatbot that helps to practice conversation in a foreign language. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 395-405). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W17-5547

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018, June). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018* (pp. 895-906). ACM. https://doi.org/10.1145/3196709.3196735

Kerly, A., Hall, P., & Bull, S. (2007). Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems*, 20(2), 177-185. https://doi.org/10.1016/j.knosys.2006.11.014

Kim, S. H., Wise, J., & Hillsley, M. (2005, October). Learning technical writing skills through peer review: use of calibrated peer review™, in unit operation lab. In *AIChE Annual Meeting, Conference Proceedings*.

King, A. (1994). Guiding Knowledge Construction in the Classroom: Effects of Teaching Children How to Question and How to Explain. *American Educational Research Journal*, *31*(2), 338-368. https://doi.org/10.3102/00028312031002338

Kulkarni, C., Kotturi, Y., Bernstein, M. S., & Klemmer, S. (2016). Designing Scalable and Sustainable Peer Interactions Online. In *Design Thinking Research* (pp. 237-273). Springer International Publishing. https://doi.org/10.1007/978-3-319-40382-3_14

Lam, Y. W., Hew, K. F. T., & Chiu, K. F. (2018). Improving argumentative writing: Effects of a blended learning approach and gamification. *Language learning & technology*, *22*(1), 97-118.

Lan, Y. F. & Lin, P. C. (2011). Evaluation and improvement of student's question-posing ability in a web-based learning environment. *Australasian Journal of Educational Technology*, *27*(4), 581-599. https://doi.org/10.14742/ajet.939

Land, S., Hannafin, M., & Oliver, K. (2012). Student-centered learning environments: Foundations, assumptions and design (Chapter 1). In D. Jonassen and S. Land, *Theoretical Foundations of Learning Environments* (2nd Edition). London: Routledge.

Liang, J. C., & Tsai, C. C. (2010). Learning through science writing via online peer assessment in a college biology course. *The Internet and Higher Education*, *13*(4), 242-247. https://doi.org/10.1016/j.iheduc.2010.04.004

Lin, M. P. C. & Chang, D. (2020). Enhancing post-secondary writers' writing skills with a chatbot: A mixed-method classroom study. *Journal of Educational Technology & Society*, *23*(1), 78-92.

Lu, J., & Law, N. (2012). Online peer assessment: effects of cognitive and affective feedback. *Instructional Science*, *40*(2), 257-275. https://doi.org/10.1007/s11251-011-9177-2

Luger, E., & Sellen, A. (2016, May). Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286-5297). ACM. https://doi.org/10.1145/2858036.2858288

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology*, *106*(4), 901-908. http://dx.doi.org/10.1037/a0037123

Macdonald, J. (2001). Exploiting online interactivity to enhance assignment development and feedback in distance education. *Open Learning: The Journal of Open, Distance and e-Learning*, *16*(2), 179-189. https://doi.org/10.1080/02680510120050334

Marzouk, Z., Rakovic, M., Liaqat, A., Vytasek, J., Samadi, D., Stewart-Alonso, J., Ram, I., Woloshen, S., Winne, P. H., & Nesbit, J. C. (2016). What if learning analytics were based on learning science?. *Australasian Journal of Educational Technology*, *32*(6). https://doi.org/10.14742/ajet.3058

Mekni, M., Baani, Z., & Sulieman, D. (2020, January). A Smart Virtual Assistant for Students. In Proceedings of the 3rd International Conference on Applications of Intelligent Systems (pp. 1-6). https://doi.org/10.1145/3378184.3378199

Min, H. T. (2005). Training students to become successful peer reviewers. *System*, *33*(2), 293-308. https://doi.org/10.1016/j.system.2004.11.003

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of second language writing*, *15*(2), 118-141. https://doi.org/10.1016/j.jslw.2006.01.003

Moxley, J. (2013). Big Data, Learning Analytics, and Social Assessment. *Journal of Writing Assessmen*t, *6*(1), 1-10.

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education*, *10*, 98-129

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, *37*(4), 375-401. https://doi.org/10.1007/s11251-008-9053-x

Nesbit, J., Niu, H., & Liu, Q. (2019). Cognitive tools for scaffolding argumentation. In *Contemporary Technologies in Education* (pp. 97-117). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-89680-9_6

Niu, H., Sharp, J., & Nesbit, J. (2015). Teaching students argumentation with the dialectical map. In *Annual Conference of the Society for Teaching and Learning in Higher Education, Vancouver*.

O'Dowd, I. (2018). Using learning analytics to improve online formative quiz engagement. *Irish Journal of Technology Enhanced Learning*, *3*(1), 30-43. https://doi.org/10.22554/ijtel.v3i1.25

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007-2004. *National Center for Education Research*.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd Ed.). Newbury Park, CA: Sage.

Pereira, J., & Díaz, Ó. (2018, June). Chatbot dimensions that matter: Lessons from the trenches. In *International Conference on Web Engineering* (pp. 129-135). Springer, Cham. https://doi.org/10.1007/978-3-319-91662-0_9

Pham, X. L., Pham, T., Nguyen, Q. M., Nguyen, T. H., & Cao, T. T. H. (2018, November). Chatbot as an Intelligent Personal Assistant for Mobile Language Learning. In *Proceedings of the 2018 2nd International Conference on Education and E-Learning* (pp. 16-21). ACM. https://doi.org/10.1145/3291078.3291115

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks*, *16*(3), 9-20.

Prins, F. J., Sluijsmans, D. M., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education*, *11*(3), 289-303. https://doi.org/10.1007/s10459-005-3250-z

Ranavare, S. S., & Kamath, R. S. (2020). Artificial Intelligence based Chatbot for Placement Activity at College Using DialogFlow. Our Heritage, 68(30), 4806-4814.

Redfield, D., & Rousseau, E. (1981). A Meta-Analysis of Experimental Research on Teacher Questioning Behavior. *Review of Educational Research*, *51*(2), 237-245. https://doi.org/10.3102/00346543051002237

Reiners, T., Wood, L., & Bastiaens, T. (2014). Design perspective on the role of advanced bots for self-guided learning. *The International Journal of Technology, Knowledge, and Society*, *9*(4), 187-199.

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, *1*(3), 181-210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Savery, J. (2006). Overview of Problem-based Learning: Definitions and Distinctions. *Interdisciplinary Journal of Problem-based Learning*, *1*(1), 9-20. https://doi.org/10.7771/1541-5015.1002

Sarosa, M., Kusumawardani, M., Suyono, A., & Wijaya, M. H. (2020, January). Developing a social media-based Chatbot for English learning. In IOP Conference Series: Materials Science and Engineering (Vol. 732, No. 1, p. 012074). IOP Publishing. Doi:10.1088/1757-899X/732/1/012074

Schunn, C., Godley, A., & DeMartino, S. (2016). The reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent & Adult Literacy*, *60*(1), 13-23. https://doi.org/10.1002/jaal.525

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review*, *1*(3), 357-375. https://doi.org/10.3758/BF03213977

Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., Kim, T., Pieper, M., Chandar, S., Ke, N. R., Rajeshwar, S., de Brebisson, A., Sotelo, J. M. R., Suhubdy, D., Michalski, V., Nguyen, A., Pineau, J., & Bengio, Y. (2017). A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Silverman, D. (2019). From the Lab to the Classroom: Team Behind Dialectical Map Encourages Critical Thinking Across Disciplines. Retrieved from https://www.sfu.ca/education-research-hub/blog/dialectical-map.html.

Sinha, S., Basak, S., Dey, Y., & Mondal, A. (2020). An Educational Chatbot for Answering Queries. In Emerging Technology in Modelling and Graphics (pp. 55-60). Springer, Singapore. https://doi.org/10.1007/978-981-13-7403-6_7

Sluijsmans, D. M., Brand-Gruwel, S., & van Merriënboer, J. J. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education*, *27*(5), 443-454. https://doi.org/10.1080/0260293022000009311

Song, D., Oh, E. Y., & Rice, M. (2017, July). Interacting with a conversational agent system for educational purposes in online courses. In *2017 10th international conference on human system interactions (HSI)* (pp. 78-82). IEEE. https://doi.org/10.1109/HSI.2017.8005002

Spiro, R. J., Feltovich, P. J., Feltovich, Paul, L., Jacobson, M. J., & Coulson, R. L. (1991). Cognitive Flexibility, Constructivism, and Hypertext: Random Access Instruction for Advanced Knowledge Acquisition in Ill-Structured Domains. *Educational Technology*, *31*(5), 24-33

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & evaluation in higher education*, 25(2), 149-169. https://doi.org/10.1080/713611428

Torrance, M., Thomas, G. V., & Robinson, E. J. (1994). The writing strategies of graduate research students in the social sciences. *Higher education*, *27*(3), 379-392. https://doi.org/10.1007/BF03179901

Wang, Y. F., & Petrina, S. (2013). Using learning analytics to understand the design of an intelligent language tutor–Chatbot Lucy. *International Journal of Advanced Computer Science and Applications*, *4*(11), 124-134. https://dx.doi.org/10.14569/IJACSA.2013.041117

Weber, R. P. (1990). *Basic content analysis (2nd Ed.)*. Newbury Park, CA: Sage.

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45. https://doi.org/10.1145/365153.365168

Xu, D., & Wang, H. (2006). Intelligent agent supported personalization for virtual learning environments. *Decision Support Systems*, *42*(2), 825-843. https://doi.org/10.1016/j.dss.2005.05.033

Yang, Y. F. (2010). Students' reflection on online self-correction and peer review to improve writing. *Computers & Education*, *55*(3), 1202-1210. https://doi.org/10.1016/j.compedu.2010.05.017

Appendix A.

Peer Review Sheet

Peer Assessment on Dialectical Map

A. Positive Comments: On which part did the writer do a good job? and Why?

B. Negative Comments: On which part did the writer need to improve? and Why?

C. Constructive suggestions: What should the writer have done in order to improve the map?

D. Pose a question: Provide one question to the writer to improve this map.

Appendix B.

Questionnaire regarding student experience with the chatbot

1. Student ID: _____
2. Name: _____
3. Email: _____
4. Major: _____
5. Academic residency: _____
6. EAL: _____
7. In the past 6 months, which of these forms of writing have you engaged in? (Please select all that apply):
    a. Plans and notes (taking notes in class)
    b. Reports and assigned work (writing emails, cover letters for jobs)
    c. Writing for publication (e.g. writing a book or blog posts)
    d. Research term paper within a course (e.g. Literature review for PHIL 120; persuasive paper for FALx99)
    e. High school essays or provincial exams (e.g. English 12 or equivalent)
    f. Other (please specify): _____
8. When you write, what strategies do you always adopt or use. (choose up to 3):
    a. Brainstorming

b. Taking notes from research sources

c. Mindmapping

d. Ordering notes

e. Making an outline

f. Drafting

g. Revising

h. Sharing ideas with a friend and receiving feedback

i. Other

9. When you revise your paper, what are your goals? (choose up to 3)

    a. Improving clarity

    b. Improving style

    c. Developing content

    d. Correcting errors

    e. Rearranging the text

    f. Reducing length

10. Generally speaking, at what point do you like to start writing?

11. The bot DD helps me identify new issues with my peer's dialectical map (Y/N)

12. The bot DD helps improve my feedback for my peer's dialectical map (Y/N)

13. What specifically did DD help me? (Open-ended)

14. When I worked with DD, he helped me change the quality of my review (Y/N)

15. Tell me how did (or did not) your review change because you engaged with the bot (Open-ended)

16. Working with DD the Peer Review/Thesis Bot was enjoyable (Y/N)

17. Working with DD taught me how to be a better reviewer of my peer's work (Y/N)

18. Did you find working with DD the Peer Review Bot helped you resolve some confusions from the instructions? If yes, what specifically was resolved? (Open-ended)

19. Describe the experience you have had in the peer-review/thesis chatbot (Open-ended)

20. What did you like about the DD chatbot? (Open-ended)

21. What suggestions would you provide to make the chatbot more effective? (Open-ended)

22. Did you have any experience with peer review before? (Open-ended)

Appendix C.

Students' comments on chatbot DD

Positive and interactive experiences:

"Positive experience and greatly assisted in the writing of my thesis statement and peer feedback."

"I enjoyed how interactive it was, rather than just having DD let you know what is, it asked students review questions about the topic afterwards to get them on track."

"Overall, working with DD was a positive experience. I felt like I came away from it with a better understanding of how to make my thesis."

"I like the pictures and the dog theme. I also liked the questions DD asked because they made me think more."

"I liked the dog and how it was very specific in helping the students learn about new unfamiliar areas."

"I like the pictures of the dog also that it gives you the choice to ask for more clarification or to skip it and continue with the lesson."

"I think it was a fun, interactive way to improve our writing. It was something unique that I had never tried before which caught my attention!"

"It was a completely new kind of activity to me...it was kind of fun (improving my work but it wasn't stressful or anything)."

"I like the dog pictures."

"it was easy to use, having options to choose from, informational with examples provided, and I liked the cute dog."

"very enjoyable because edit made it wayyyy more enjoying very helpful as a start, but maybe not for specific questions."

“really fun and organized, also gave me a really clear start for my thesis statement and made me realize how much I need to change the specific ness of my statement.”

“it was enjoyable and simple to use & very helpful.”

“quick to respond and explains everything if you don't understand.”

“It was fun and talking to a dog made it a more enjoyable experience.”

“It was modelled as if you were chatting to a dog which was amusing. It gave clear instructions and advice on how to make a thesis/feedback more effective and higher quality while ensuring the student did the work.”

“Chat was easy navigate - dd bot provided good feedback and explanation.”

“it was simple to use (point and click).”

“I can answer by clicking a button instead of typing the answer.”

“user- friendly - cute (doggo).”

“promotes students to be engaged.”

“It was enjoyable I learned the best way to give feedback; It was interactive and the dog photos.”

“It was a relatively positive experience. I enjoyed working with the bot and it helped me create a much better thesis statement. I liked that DD was supportive, and I liked the dog photos!”

Deepen reflective thinking:

“It gave me time to think and reflect on thesis statement and arguments it helped me evaluate other arguments and create an expectation for my paper.”

“Promotes students to be critical and reflective in their opinions and ideas - self regulated learning - beneficial when organizing thoughts and ideas in a coherent way.”

Resources and instructions guidance:

“It provides good information for students who have no idea what to do.”

“clarified the instructions.”

“DD was easy to use with clear instructions.”

“It was quite detailed in explanation of each section/topic.”

“After submitting my writing to the chatbot, it provided a breakdown of an effective review or thesis statement and allowed me to compare mine to the rubrics.”

“The emphasis on organization that provided clear instructions and an easy-to-follow format.”

“Learning specific types of reviews/comments.”

“Detail specific feedback is beneficial.”

Overall experience:

“Overall, it was helpful.”

“I like the concept of the DD chat bot, and I think it would be helpful for students.”

Suggestions:

“Perhaps having the DD bot on a more accessible website that can be accessed more readily and quickly would be more effective.”

“Maybe try to condense replies/answers so it is not too long and try incorporating some self-tests/quizzes! That would be fun.”

“Ask more open-ended and specific questions”

“More examples provided for peer-review commentary”

“Being able to provide more feedback”

“Not enough time to evaluate everything - when doing the peer review, it kept taking the conversation back to additional resources and not moving on to the actual evaluation.”

“I would have a mandatory the peer review to help with my own writing instead of using the sample. Also I think it would be more helpful to have the dialectical map completed before using the chatbot as a way to check over writing.”

“The system itself is wonderful, but maybe the graphics could be better.”

“Updating the program so it responds faster.”

“Faster pace, not so much back and forth, get to the point of the activity.”

“Faster response times and more options for potential questions.”

“The response time could be little faster.”

“Improve speed of response, provide more options and interaction for the providing feedback chatbot.”

“It was okay, the program is slow and I don't it is as useful as talking to a TA or professor.”

“Wait time was kind of long.”

“I felt that the structure of DD was very predictable, meaning that there was an order of points that DD made, and I felt that for people who did not need help with certain aspects of the thesis did not need to go through the steps. Maybe if DD had the ability to give feedback, it would be more productive so a student knows exactly what to work on.”