Chatbot learning partners: Connecting learning experiences, interest and competence.

AUTHORS:

Luke Fryer, Hong Kong University, Centre for the Enhancement of Teaching and Learning, lukefryer@yahoo.com
CPD 173, Pokfulam Rd.,The University of Hong Kong, Hong Kong (Contact author)

Kaori Nakao, Seinan Gakuin University, Nakao.K.N@gmail.com

Andrew Thompson, Kyushu Sangyo University, LERC, thompson@ip.kyusan-u.ac.jp

The authors declare that they have no conflict of interest.

This research was not supported by a research grant.

Acknowledgements: We would like to acknowledge Aaron Gibson and Zelinda Sherlock for their aid in collecting data utilised in the current study. We would also like to acknowledge Dr. Alex Shum for his very helpful review of a previous version of this manuscript.

The Official Online Location is:

Please reference this papers as:

Fryer, L. K., Nakao, K., Thompson, A. (online). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior.* doi 10.1016/j.chb.2018.12.023

Chatbot learning partners: Connecting learning experiences, interest and competence.

1 Introduction

Technology has and will continue to have a dramatic effect on teaching and learning. Perhaps more than any other aspect of education, advances in technology have opened doors for language-learners. From cassette tapes and VCRs, to CD/DVDs and MP3/4s, to VOIP telephony and the plethora of resources the Internet now provides, technology has made it easier for second language learners, and increasingly possible for students in foreign language learning contexts to improve their language skills.

If technology were a tent pole raising the pavilion of language learning, then its sharp end would be software capable of supporting language learners through intelligent, scaffolded practice. The field of chatbots seemed poised to begin this shift in how we learn or at least practice a new language. Being free and online, chatbots could provide opportunity for language learners from all parts of the globe to actively communicate in their chosen second/foreign language. In addition to supporting competence development, chatbots might also have a role in supporting the kind of motivation students need to persist in language learning (Fryer & Carpenter, 2006). While the promise of chatbot technology for the language learning community has not yet been fully realised, their presence on the Internet continues to grow faster than ever (Dale, 2016). Even in their as yet under-developed state, chatbots present a free and ubiquitous source of language interaction for many students learning English as a Foreign Language (i.e., in countries

where English is neither the country's first or second language). It is therefore essential that research strive to understand how these potential language partners might be put to use best, both inside and outside formal education.

Seeking to address this general aim, the current research extends an experimental study comparing university students' interest in chatbot vs human partners and the effect of these partners on students' interest in their language course (Fryer, Ainley, Thompson, Gibson & Sherlock, 2017). The current study was undertaken with the same group of students, during the next semester of the same course. The current study builds directly on Fryer et al., (2017) by retesting the same students' interest in the (same) chatbot vs. a random human language partner. This extension facilitates longitudinal tests seeking to explain students' interest in these two language partners months later. Finally, the present follow-up study adds a qualitative element to our understanding by examining the merits and demerits of a chatbot/human partner from the perspective of language-learning students across a range of language proficiencies.

Through the research described, the current study aimed to examine the complex interaction between the strengths and weaknesses of current chatbot technology, language learners' language ability and their interest in continuing to engage with the software. Results from this study are targeted at providing direction for educators in using the wide variety of chatbots currently available online, and supporting developers in beginning to fill this sure-to-grow niche in "casual" language learning technology.

The current study, along with the experimental study it builds on (Fryer et al., 2017), narrows the focus of Fryer and Carpenter's (2006) examination of the motivational implications of chatbots, and educational technology more broadly, for

language learners. The interest students experienced when conversing with a chatbot and human partner were compared to provide a balanced perspective on how chatbots, that the internet is already teeming with, might be increasingly effective language learning partners.

2 Literature review

2.1 Chatbot development

Chatbots (Chatterbots) began as a computer-based experiment with language. Joseph Weizenbaum's ELIZA (1966) is a famous early attempt at creating software which could maintain a conversation with a human. The ELIZA program was designed to replicate the kind of questioning interaction a psychoanalyst might utilise and thereby engage an individual in discussion. Attesting to the potential of this early chatbot—and the many chatbots that have and continue to follow—is the fact that despite the early chatbot's weaknesses (limited to very narrow ranges of questions), users have reported preferring to discuss their feelings with machines rather than other humans (Block, 1981). There is also the enduring popularity of Eliza, which is still used online (Heller et al., 2005) decades after its inception. Both observations are as true for early chatbot interaction as it is today, with researchers highlighting the persistent communication with chatbots many individuals choose to undertake (Hill, Ford, & Farreras, 2015).

The intervening five decades of chatbot development have seen steady, if only small improvements to chatbot technology. For the second half of this period, the annual competition for the Loebner prize presents perhaps the clearest trajectory of this development. The Loebner competition is an application of Turing's famous test, a test

no chatbot has yet to pass. A survey of the Loebner competition's winners (Bradeško & Mladenić, 2013) suggests that chatbots have developed from simple pattern matching systems like Weizenbaum's pioneer effort to steadily becoming increasingly complicated in their patterns of interaction and computer reasoning; however, no substantial new breakthroughs have recently emerged.

Since ELIZA's early beginning, chatbots have seen increasing use across the Internet. Chatbots have played a wide range of roles within commerce (for practical recent examples see Huang et al., 2014; Lasek & Jessa, 2013). In addition to their ubiquitous potential for sales, the prospects for chatbots within fields as far ranging as stress management (Huang et al., 2015) and library support (Allison, 2013) are constantly being tested. The most natural and potentially powerful application of chatbots, however, is in line with their fundamental nature: language practice.

2.2 Chatbots and language learning

Computer Assisted Instruction (for a recent review see Voogt, Fisser, & Wright, 2015) has a substantial history in the area of human-technology interaction supporting learning. Considerable development and research have focused on the potential of software based intelligent tutors or pedagogical agents (for an early review see, Burns & Capps, 1988). Mayer and colleagues have conducted much of the pivotal research in the area of pedagogical agents, making important advances in areas such as the role of animation (e.g., Moreno, Mayer, Spires, & Lester, 2001) and supporting text (e.g., Moreno & Mayer, 2002). The idea of language tutors has since been pursued to support language learning in classrooms (e.g., Graesser, Chipman, Haynes, & Olney, 2005) and increasingly with mobile technology (e.g, Kukulska-Hulme & Shield, 2008).

A second area of development has come from outside of both education and language learning circles: Chatbots are programs developed to engage in conversations with humans. Their role within formal language education is currently tangential at best. What little research there is in this area has generally focused on their comprehensibility and the motivation they inspire in their users. One of the earliest examinations of chatbot applications for language learning suggested that language-learning students generally enjoyed conversing with a chatbot (Fryer & Carpenter, 2006) and that some students preferred conversing with the chatbot over other students/teachers. At that stage, chatbots were suggested as being primarily useful for motivated and/or advanced students. Studies that followed examined the role of chatbots in supporting motivational elements such as learner autonomy (e.g., Shawar & Atwell, 2007), intrinsic motivation (e.g., Jia & Chen, 2008) and an inquiry-orientated frame of mind (Goda, Yamada, Matsukawa, Hata, & Yasunami, 2014).

Early research evaluating chatbots as language practice tools indicated their limitations (Coniam, 2008; Fryer & Nakao, 2009). Both user input (e.g., necessity of correct spelling) and chatbot output (e.g., inability to stay "on topic") concerns were raised as issues that needed to be overcome for chatbots to be of widespread usefulness to language learners. Marking the relatively small improvements in chatbot language competency, few educational researchers have assessed chatbots language competence again until recently. Coniam (2014) evaluated five well-known chatbots concluding that chatbots have significantly improved, with three of the five presenting 90% of their answers as being grammatically acceptable. While this evidence was hopeful, it failed to push chatbots into the clear category of broadly useful for language practice. In contrast,

Hill et al. (2015) analysed 100 messaging conversations and found that humans carried on significantly longer messaging conversations with the chatbot than with other humans. While each message sent to the chatbot was shorter and the vocabulary not as rich, this finding demonstrated the ease with which humans can communicate with chatbots and the quantity of conversational engagement that was possible.

Seeking a concrete measure of improvement over the original Weizenbaum chatbot, Shah, Warwick, Vallverdú, and Wu (2016) carried out an experiment directly comparing five established modern chatbots with the original ELIZA chatbot. A comparison of the scoring of the quality of conversations with the modern versus the original ELIZA resulted in conversations with the modern chatbots being of significantly higher quality across a range of areas. At the same time, however, these improved chatbots were often vague and on occasion misled the human participants.

As highlighted, chatbots have improved during the past five decades and despite lacking a major breakthrough in their language interaction skills, humans still find them intriguing enough to play with. This, in and of itself, is hopeful, because it is well established that play is an essential component of the learning process (e.g., Piaget, 1976; Vygotsky, 1978). Through play we can test ourselves and reach beyond our current abilities under non-threatening conditions. These are affordances which technology has an established strength in providing (e.g., Roussou, 2004). The potential role for chatbots to inspire curiosity and "in the moment interest" therefore deserves further examination.

2.3 The development and correlates of interest

Current conceptions of the psychology of interest present a robust developmental theory (i.e., four-phase model of interest development; Hidi & Renninger, 2006;

Renninger & Hidi, 2011) for researching the motivational potential of chatbots for language learners. Consistent with the four-phase model, it is generally understood that an individual's interest can be divided into situational (largely affective and transient) and individual (increasingly inclusive of value and epistemological components) interest. Also, interest is understood to be content specific (i.e., specific to one domain, rather than generalised across many domains). Standing on these agreed general theoretical foundations, the four-phase model of interest development describes an individual's interest, beginning with initial situational experiences of fun, curiosity and stimulation (triggered situational interest; phase one). Then, with increasing experience, an individual's interest can become more sustainable (maintained situational interest; phase two). Over time and with increasing knowledge of a topic and an increasingly clear sense of value for the area, this interest can potentially develop into the first stage of individual interest (emerging individual interest; phase three). Finally, an enduring form of interest, marked by a consistent desire to reengage with the topic can develop and become a standing source of motivation for a specific domain of study (well-developed individual interest; phase four).

The four-phase model is a useful means of understanding interest and its development, and is particularly relevant for attempting to explain the potential role of chatbots within language learning; this is in part, because of its collative and clear developmental organisation. By collative, we mean interest, depending on its stage of development, can be made up of a number of constructs (value, enjoyment, knowledge, and a desire to learn more). By developmental we refer to the four stages described by Hidi and Renninger (2006, 2011). The developmental nature of the theory enables

researchers to more accurately describe how interaction with a chatbot might lead to short and long-term effort, which are key correlates of interest (Tobias, 1995). Employing such a model, and undertaking a careful analysis of the interest that students experience interacting with chatbots, might help us shape them as better tools for language learners.

2.4 Modelling Interest development within formal education

The four-phase model (Hidi & Renninger, 2006; Renninger & Hidi, 2011), while a powerful tool for organising our understanding of general interest development, unfortunately often fails to fit clearly within the environmental constraints presented by many formal classroom settings. For example, student learning is often highly structured in most contexts, with little opportunity , for the pursuit of students' budding interests. From early education through to the majority of tertiary experiences, formal education consists chiefly of specific subjects, which students take year or semester-long courses in, within which students undertake a wide variety of tasks/activities.

Fryer, Ainley and Thompson (2016) sought to model interest development in this highly-structured environment. This model was built firmly on the principles of the four-phase model (i.e., assessing a collative conception of interest and acknowledging the developmental nature of interest). At the same time, it also aimed to hypothesise about, and test connections between students' on-task experiences, their interest in their current course and their developing interest in the domain of study. In an initial longitudinal study employing Structural Equation Modelling, Fryer et al. (2016) observed complete mediation of students' interest in tasks (i.e., group vocabulary review exercises) through their future interest in the course, to their general interest in studying the English language at the end of the course (domain-level interest). Furthermore, the role of self-

efficacy in supporting task interest varied depending on initial academic self-concept and how many times they undertook the activity (different content, but same activity each time). Research findings supported a complex relationship between interest and both perceived and actual competence (i.e., self-efficacy, self-concept and standardised assessment of language fluency; Fryer & Ainley, 2018; Fryer et al., 2016; Fryer, 2015; Hidi & Ainley, 2008; Silvia, 2003). These findings point toward the important role of students' competencies (actual and perceived) within their interest at both the domain (English in general) and task (specific activities for learning English) level.

2.5 The current programme of research and the current study

The present study seeks to examine chatbots as language practice tools, specifically from the perspective of their potential role in stimulating and then supporting an individual's interest in language learning. The current project builds on and extends a recent experimental study (Fryer et al., 2017), testing the developmental nature of students' interest in a course of study across three structured conversations. These conversations were conducted with both human and chatbot partners, utilising a counter-balanced design. All conversations with the chatbot were conducted with speech-to-text software for student-to-chatbot output and simple text for chatbot-to-human output.

Both the prior (Fryer et al., 2017) and current study utilised Cleverbot. Cleverbot is a recent version of a series of chatbots developed by Rollo Carpenter (NA, 2015). Cleverbot and its well-known predecessor (i.e., Jabberwacky) have seen considerable use in previous early and recent studies (Fryer & Carpenter, 2006; Hills et al., 2015). Cleverbot was found to make relatively few grammatical errors (Conaim, 2014) and was therefore selected for use as the chatbot partner in both the initial and current study.

In the initial experimental study (Fryer et al., 2017) students' interest in tasks was measured via the completion of a short survey immediately after each conversation (human-human and human-chatbot). Structural Equation Modelling and repeated ANOVA of the resulting longitudinal data (Times 1, 1.5, 2, and 3.5; see Figure 1) revealed two findings of theoretical significance. First, controlling for prior interest in the course, only students' interest in the human conversations predicted increased interest in the course.

Second, a "novelty effect" (a longstanding and still important issue for educational technology; see Chen et al., 2016; Clark, 1983) for the conversations with the chatbot was observed. For the first conversations with the human and chatbot partners, no statistically significant difference between students' initial interest in the two conversation partners were found. Between conversations one and two, students' interest in conversing with the chatbot dropped significantly; interest in conversing with a human partner, however, remained consistent. These results were not a good sign for the potential usefulness of chatbots for stimulating meaningful interest in language learning. However, the longstanding (Fryer & Carpenter, 2006) and recent (Hill et al., 2015) empirical evidence pointing towards sustained human interest in talking to chatbots suggests that a more fine-grained examination is necessary. Given the essential role of students' prior language competence, such an examination must take students' language skills into consideration. Finally, much like the Turing Test itself, an essential benchmark for understanding chatbot-human interaction is human-human interaction. Therefore, any meaningful test in this area must make such comparisons possible.

The current study acknowledges the essential role of task interest for the development of interest in the English language at the domain level, mediated by course interest (Fryer et al., 2016). The current study extends and elaborates on experimental examination in Fryer et al. (2017) of the short-term novelty effects and connection between interest in task learning partners and interest in a course. The current study therefore pursued a mixed-methods (qual/QUAN) extension by adding an additional measurement of students' interest in the same two language practice tasks (human-human vs. human-chatbot) and collecting supplementary open-ended textual feedback. This replication and qualitative data were collected four months after the end of the original study, during the subsequent semester with the same students.

Modelling in the current study included standardised term test results and self-reported interest in tasks, course and the general domain of study. The students' performance on a standardised listening/reading semester-end test (the institutional measure for language proficiency has been employed successfully in past studies: Fryer et al., 2016; Fryer, 2015). As well as an additional measure of task interest, the current modelling included six additional self-reported measurements of interest from the previous semester. Four of these were from the experimental study (Fryer et al., 2017) we aimed to extend and elaborate on, as well as two more measures (Time 0 domain interest and Time 3 task interest). Please see Figure 1 for a complete breakdown of all measures and the division between the previous experimental study and the current study.

Through the use of both quantitative and qualitative data, the current study aimed to provide a comprehensive examination of students' interest experiences. This study aimed to integrate students' perceptions of the merits and demerits of chatbots as

language practice tools into this examination. In the current study, merits and demerits refers to students' perceptions of the qualities of the human or chatbot partners which supported, failed to support or hindered the partner from being effective during a 15-minute language practice task (based on weekly classroom materials) with the student.

==========FIGURE 1 ABOUT HERE==========

3 Research questions

The current study was organised around three general research questions:

1) Did the decreased interest (Time 1 to Time 2 and persisting to Time 3; see Figure 1) in conversing with a chatbot (observed in Fryer et al., 2017) persist across a 20-week gap to Time 5 (Research Question 1)?

2) What was the predictive relationship between prior task, course interest and language (listening) competency (Time 1 to 4) for task interest at Time 5 (Research Question 2)?

3) How do students' prior competency, current speaking task interest, overlap as clustered within the perceived merits and then demerits of the chatbots as a source of conversation practice (Research Question 3).

4 Methodology

4.1 Participants

Of the original students from the prior experimental study (n =122; Fryer et al., 2017) 91 students (22 Female) participated in the current study. Students were in their first and second year (between 18 and 20 years of age) at a private university in Western Japan. Participating students came from five of the university's seven faculties (Engineering, Management, International Studies, Fine Arts and Economics) and were studying within a coordinated compulsory English as a foreign language program. As with the previous study, all students were placed in their English language classes based on previous performance on a standardised listening/reading test (see Stewart, Gibson & Fryer, 2012). Specifically, as the current class focused on oral communication, students were placed in classes based on their listening test results. All students participating in the study were in the intermediate banding of listening proficiency (i.e., having basic communicative fluency in day-to-day English language). Within this program, textbooks, studying material (Fryer, Anderson, Stewart, Bovee & Gibson, 2010), e-learning (Bovee & Fryer, 2011) and pre/post vocabulary tests (Stewart, Fryer & Gibson, 2013) were coordinated across all classes. This ensured that students received a consistent learning experience regardless of their assigned class.

4.2 Procedure

One week after the initial experimental study (Fryer et al., 2017) and 14 weeks prior to the current extension study, (see Figure 1 for a detailed presentation of the prior and current research schedule) students sat a 60-minute standardised listening/reading competency test. Fifteen weeks after the experimental study, students read an outline of

the current study and chose whether to have their self-reported and achievement data included in the current research output. Students' anonymity was guaranteed and no inducement was given to encourage students to participate in the research. Ethical permission to undertake the study was granted from the educational centre where the study took place.

The structured conversation used for the current study (same materials and methods as with the prior experiment; Fryer et al., 2017) was developed from the students' course materials (the textbook all participating classes used). The students were randomly organised into two groups. One half of the students spoke with a randomised human partner for 15 minutes employing the supplied structured dialogue (explained and handed out to students at the beginning of the class) and the other half spoke with the chatbot on a Nexus nine-inch tablet. Consistent with the research this study builds on (Fryer et al., 2017), the structured dialogues were developed directly from textbook materials students were currently learning. The dialogue provided 4-5 minutes of dialogue they could read to each other and then additional content to support students in continuing to converse (e.g., questions, questions stems, topic and reply stems). Again, consistent with Fryer et al. (2017), the chatbot conversation was undertaken employing the tablets' native Google Chrome browser speech-to-text software for students' spoken output. Textual output for the chatbot was utilised. The same structured dialogue was used for the human-human conversation. Students switched conditions after 15 minutes to the other partner.

Consistent with the original experiment, students' interest in the speaking task was assessed immediately after the human-human and human-chatbot conversations with a five-question six-level Likert scale examining task interest (the same as the

previous study). After both conversations were completed, students were asked to provide short answer feedback regarding the relative merits and demerits of both chatbot and human-partnered tasks.

4.3 Instrumentation

In both the prior and current study, three different short Likert-based (one to six, from totally unlike me to totally like me) interest surveys were included for interest measurement. In addition, an open-ended survey regarding the merit/demerits of human and chatbot conversations and a 180-item listening and reading competence test were also instruments for the current study.

The domain interest survey (from Ichihara & Arai, 2004) consisted of three items (e.g., "I find English interesting" and "English arouses my curiosity"), the course interest survey consisted of four items (e.g., "I am fully focused on learning English in this course" and "This English course is interesting") and the task interest survey consisted of five items (e.g., "This activity is personally meaningful" and "I enjoyed learning English in this activity"). The course and task scales were developed for Fryer et al. (2016) and employed within Fryer et al., (2017). For both scales, a pool of theoretically relevant items was developed based on the existing interest literature, with split-half EFA and CFA conducted on the pilot data to resolve on a parsimonious set of items for task and course interest scales.

The listening/reading test was an institutional, standardised measure of students' English language skills (used successfully in a previous interest-based study; Fryer et al., 2016; Fryer, 2015) and took approximately 60 minutes to complete. Only the 90 listening focused items from the test were used in the current study because listening improvement

was a course goal for the classes included in the current study. Furthermore, their listening achievement was used to place them in their current class and contributed to 20% of their semester-end grade.

Finally, the open-ended survey instructed students to write up to three merits and demerits for each the human and then chatbot partners. The questions and students' answers were in Japanese. If students did not perceive any merits/demerits, they were asked to simply write "none". The instructions stated that their answers should be in Japanese and ranked in terms of strength (i.e., strongest to weakest).

5 Data collection and analysis

Figure 1 presents a detailed schedule for the data collection across both the prior and current study. All Likert surveys were undertaken online with the tablets used for the chatbot conversations. The open-ended survey was completed on paper after the tablet survey.

All quantitative analyses for the current paper were undertaken with JMP 9.01 (SAS, 2007-2011). Missing quantitative data (<1%) were imputed with a Robust Maximum Likelihood Estimator prior to analyses being conducted. Cronbach's Alpha was calculated to estimate the reliability of the scales used. Calculating the pairwise correlations for all constructs followed to provide a general overview of the entire data set.

To address Research Question 1, a *t*-test was undertaken with the original measure of task interest (chatbot Time 2) and the current measure of task interest (Time 5), which exhibited a drop in task interest interpreted as a novelty effect. This was done to estimate the durability of the novelty effect found in the prior study (Fryer et al., 2017).

Seeking to estimate the relative predictive importance of prior variables for future task interest, regression analysis was undertaken (Research Question 2). While latent Structural Equation Modelling is perhaps the best tool for answering this type of question, the relatively small sample size prevented its use. Instead, an exploratory multiple regression analysis was undertaken to assess the variance explanation of prior competency and interest (task, course and domain) for future communication tasks with human and chatbot partners. Despite its weaknesses, multiple regression (Forced Entry) is still a useful tool for gauging the relative importance of variables (for an overview of it strengths and weaknesses see Montgomery, Peck & Vining, 2015; for a review of its continued relevance see Whittingham, et al. 2006) at two stages: 1) a full model with all prior interest measurements and prior competency tested together for their predictive power, 2) a model with only significant predictors from the full model. This exploratory approach was undertaken due to the lack of prior research for the construction of clear hypotheses regarding the prediction of conversation task interest. The second model (significant predictor model) was conducted, exclusively to suggest direction for future confirmatory tests of these questions. We ran supplementary regressions each with one independent variable to test the predictive validity of students' interest in the course (Time 3.5) and their initial interest in talking to a chatbot or a human (Time 1).

Coding and content analysis of students' open-ended feedback regarding the merits/demerits of the human and chatbot task were undertaken by two coders, a bilingual Japanese and English native speaker, on the Japanese text rather than on translations of students' answers (Research Question 3). Coding began with independent development of codes for the data by researchers. Comparison and discussion resulted in a list of codes

for each of the human and chatbot merit/demerit open responses. Two researchers then coded the entire data set independently, compared their results and resolved differences through discussion (see Brislin, 1980). Employing these codes for students' experienced merits and demerits of the chatbot task engagement, Z-scores of students' prior competency and task interest (for the chatbot partner) were organised based on these two layers of student feedback. First, the prior language competency and chatbot task interest data were organised by chatbot task demerit and then at a second, broader level within chatbot task merits. This was done to provide insight into the balance of these perceptions relative to prior language skills and students current task interest in the chatbot conversation partner (Research Question 3).

6 Results

As background for the current study, reliability, descriptive statistics and reliability for all scales were calculated and examined (see Appendices Table 1). All scales presented good reliability (> .70; Devellis, 2012) and pairwise correlations were < .90, which is considered the point beyond after which multicollinearity issues need to be considered (Tabachnick & Fidell, 2007).

*6.1 Regression model and novelty t-*test

The regression tests were undertaken to test the predictive power of prior interest (task, course and domain) and language competence for human-human and human-chatbot language practice. Predictive modelling (Table 1) resulted in substantial variance explained for interest in both the human (full model R^2 = .67, final significant predictor model R^2 = .66) and chatbot (full model R^2 = .65, final significant predictor model R^2

= .61) conversation. Prior interest in the course explained the same amount of interest in both conversation types (R^2 = .49). Prior interest in conversations with human partners explained the bulk of the variance in both Time 5 measures of interest (chatbots R^2 = .51; human R^2 = .47). Table 2 presents a clear breakdown of the variance explained.

==========TABLE 1==========

A dependent *t*-test for Time 2 and Time 5 chatbot conversation interest established that students' interest in conversing with the chatbot had a significant, moderate rebound (t = 1.82, df = 90, Std error = .11, $p < .05$, Cohen's d = .35). The rebound in interest did not, however, bring it back to its highest level at Time 1.

6.2 Merit/Demerit coding

Independent coding (by two coders) of students' open responses to the four questions, followed by discussion resulted in six codes for chatbot partner merits:

1 Convenient: e.g., "It (chatbot) was convenient."; "It (chatbot) was quick to use."

2 Learn-more: e.g., "It is good for pronunciation practice."; "The chatbot asked me lots of questions, so I could get a lot of conversation practice."

3 Task interest: e.g., "The (chatbot's) answers were interesting! Fun!"; "It (chatbot) was more interesting than usual class."

4 Technical benefits: e.g., "When I could not pronounce correctly, it (chatbot) will not be able to read it. That is a good point."; "It is easy to use the chatbot input and output."

5 Value: e.g., "It (chatbot) is useful for learning grammar."; "It (chatbot) is useful for pronunciation."

6 None.

Four codes for chatbots' demerits were resolved:

1 Ability problem: e.g., "It is too difficult."; "I feel too rushed and it is too complicated."

2 Communication problem: e.g., "We could not communicate."; "It gives me nonsense answers."

3 Technical difficulty: e.g., "Many times it cannot understand my voice."; "It (chatbot) does the wrong things sometimes."

4 None.

For the human merits five codes were resolved:

1 Communication ease: e.g., "I can just ask them anything I want."; "I can communicate with them easily."

2 Learn-more: e.g., "I can improve my conversation skills."; "I can improve my pronunciation."

3 Social benefits: e.g., "I can make friends with people I do not know."; "I get to talk to people I don't know."

4 Task interest: e.g., "It is fun to talk in English together."; "I enjoyed it.";

5 None.

Human partners demerits were resolved into seven codes:

1 Communication problems: e.g., "There are people who are not good at communication.", "It is difficult to talk if your partner does not open up to you."

2 Not interesting: e.g., "Not interesting."; "It was just a bother."

3 Not Learn-more: e.g., “We can’t check to see if our English sentences are correct or not if it is just us.”, “I can’t improve my English because I just talk to a partner at the same level of English as me.”

4 No value: e.g., “We always talk to partners and it takes a lot of time.”, “Not a good use of class.”

5 Social problem: e.g., “When we don’t understand, we just use Japanese.”, “I know my partners’ answers because it is always the same.”

6 Ability problems: e.g., “When I don’t know the words, the conversation comes to a dead end.”; “It is hard to think of the answers on the spot.”

7 None.

To test the substantive role of the coded categories, ANOVAs of students’ prior competency and the two concurrent interest measurements were undertaken using the coded merits and demerits as independent variables. The ANOVAs resulted in significant differences for interest in human conversations based on their coded merits, interest in chatbot conversations based on coded demerits, and prior competency on coded chatbot merits and demerits (see Appendices Table 2 for means, standard deviations, ANOVAs and Tukey’s HSD). The only result directly relevant to the current study was the important role prior competence played within chatbot-partnered relative to human-partnered experiences.

7.1 Intersection between chatbot’s qualitative merits and demerits

To provide an in-depth perspective on the interaction between students’ reported merits and demerits for conversing specifically with the chatbot (the focus of the current study), these were organised in a two-step manner, with prior competency and self-

reported interest in the chatbot nested within these two layers of perspectives on the chatbot conversation experience (perceived demerits organised within perceived merits of the chatbot partners; Figure 2).

==========Figure 2==========

Figure 2 indicates a few interesting relationships between the perceived merits/demerits of chatbots and students' interest in the chatbot task and their prior listening competency. First, technical problems were dispersed across all of the merits reported, suggesting that many students experienced difficulties using the chatbot. These technical problems were generally short-lived (e.g., internet connectivity), but could also be sustained (e.g., speech-to-text issues). The general results suggested that the intersection of the merits and demerits students perceived the chatbot as having, affected the interest experienced. For example, students who felt that the chatbot offered opportunity to Learn-more, but also had Communication Problems, experienced more than one SD greater in their task interest than students who perceived the chatbot partner as convenient or fun (situational interest) and experienced the same Communication Problems. This contrast is particularly striking for Convenience students who also scored better on the prior language competence exam than the Learn-more students. A similar comparison is clear between Learn-more/None and Technical benefit/None students' situational interest.

A clear pattern of prior competency was evident for students reporting the primary chatbot demerit as Ability Problems: these students presented universally low

prior competency. These results were to be expected and are an important source of validation for the coding of the open-ended textual feedback data.

The pattern of task interest experienced for students reporting Technical Problems (the most common demerit for chatbots) is important as it may suggest how such demerits might be overcome. Students who reported the chatbot as being convenient or not having any merits, reported task interest nearly one SD below the mean. Students who reported the chatbot as supporting more learning, having situational novelty, having technical benefits or value, however, reported task interest above the mean despite noting different problems with chatbot interaction. However, only students reporting chatbots as supporting them in "Learning-more" expressed above average interest despite experiencing communication or technical problems.

8 Discussion

The *t*-test suggested that the decreased task interest in the chatbot partner, presented by Fryer et al. (2017) as a novelty effect, does see a small, but significant rebound five months later (Research Question 1). Regression analysis was used to model the variance in the task interest in human and chatbot partners (Time 5). Results pointed to prior interest in talking to human partners first and interest in the course second as explaining substantial variance in both Time 5 chatbot and human task interest (Research Question 2).

The coding of students' perceptions of the merits and demerits of human and chatbot partners presented between four and seven codes. The students' perceptions of the chatbots' demerits organised within the chatbots' merits were used to profile students' interest in the chatbot tasks and prior competency indicated that (see Figure 2) 1) the

coding was consistent with prior competency; 2) students who saw the chatbot as helping them "Learn-more" experienced higher levels of interest in the task, compared to other merits, even when facing communication problems; 3) finally, technical problems were also experienced quite differently depending on the perceived merits of the chatbot, with the merit "Convenience" in particular not being supportive of interest in the chatbot task (Research Question 3).

Longitudinal regression modelling and difference testing across academic semesters suggested that prior interest in speaking with human partners, not chatbots, as well as students' prior ability are important correlates of students' interest in potential chatbot learning partners. The qualitative extension this study adds to Fryer et al. (2017) indicated if the strengths and weaknesses of our current chatbot technology might be understood together. Their convergence on students' interest in chatbot partners and interaction with students' prior ability suggests that students' perceptions of the chatbot as a tool for "Learning more", might mitigate some common chatbot weaknesses due to their slowly improving, but still underdeveloped communicative intelligence. The following sections will address the theoretical and practical implications of the study's findings.

8.1 Theoretical implications

Given that increasing motivation is a common impetus for using educational technology and a strength of chatbots (Fryer & Carpenter, 2006), the novelty effect observed by Fryer et al. (2017) is a source of considerable concern. Results from the current study suggest that the decreased interest in the chatbot partner (Fryer et al., 2017) does see a significant rebound given enough time. The question that remains is whether

the interest stimulated by reengaging with the chatbot contributes to interest in the students' course (and thereby the broader domain)—as it failed to do during the first term (Fryer et al., 2017). Or is it simply a source of triggered situational interest that fails to contribute to interest development as organised by the four-phase model (Hidi & Renninger, 2006; Renninger & Hidi, 2011). A continuing programme of experimental research is necessary to resolve this fundamental question about the usefulness of the Internets' burgeoning population of chatbots for supporting the development of interest in learning a foreign language.

Consistent with Fryer et al. (2017), interest in human-partnered conversations tasks were of paramount importance in predictive modelling for the current study. The important difference for the results of the current study, however, is that interest in conversations with human partners were the chief predictors of both future chatbot and human-partner task interest. These findings suggest that educators should not rely entirely on chatbots to stimulate interest in language practice, not even interest in future practice with chatbots.

The reason behind the suggested importance of human conversation for students developing interest in a language course (and thereby learning a language more generally, see Fryer et al, 2016) might be directly linked to the four-phase model's (Hidi & Renninger, 2006; Renninger & Hidi, 2011) organisation of interest development. Based on this model, for students' interest in the course to develop, their value for the course's topic, language in this case, must grow. It is reasonable to suggest that while talking to the chatbot might have been stimulating, it was unlikely to have helped students see the value of the language they were using. From the other side of the equation, it is easier to

see how talking to another student might have addressed the usefulness of the target language: i.e., "I need this language to get my point across to them."

The organisation of students' interest in the chatbot task and their prior language ability within the students' experience of the weaknesses and strengths of the chatbot partner pointed toward the perception of chatbots as a means for "Learning more" than they could with a human partner in some respects (compared to seeing the chatbot as a convenient tool) as potentially being an important moderator for supporting interest in the chatbot learning partner. The desire to learn more—to improve and to reengage with the topic—is an essential component and outcome of interest as it develops (Hidi, Renninger, & Krapp, 2004). Students who see the chatbot as a tool for helping them improve, fruitfully engage with the language in a way human partners cannot or will not support, might be short-circuiting this process and thereby supporting interest development.

8.2 Practical implications

We would suggest that there might be a range of measures that could ameliorate some of the issues raised thus far with chatbot use. Above and beyond general approaches to chatbot use (Fryer & Nakao, 2009), practical implications for teachers will be addressed, followed by implications for the design of future chatbots.

The first recommendation for teachers is that the spaced use of technology could be a preliminary means of addressing the short-term novelty benefits of educational technology such as chatbots—at least in their present form. The second is that students' prior competence needs to be taken into consideration both for current practice and future development. Teachers need to be sure that activities and overall support are sufficient to ensure all students can be meaningfully stimulated by the chatbot-human conversation

experience. Both the current study, and the prior study it builds on, strongly indicated that interest in language practice experiences chiefly arise from human-human conversation experiences. The current study's results reinforce past findings, indicating that if teachers want students to be engaged by chatbot language practice then they need to ensure students first have ample chance to be stimulated by human-human practice. Human-human conversation predicts interest in both forms of conversation tasks and interest in the course itself as well (Fryer et al., 2017; 2016).

As indicated by the current study, and with past findings (Fryer & Ainley, 2018; Fryer et al., 2017; Fryer, 2015), the competencies and motivations students bring with them can have a powerful effect on their future experiences and motivational development. Recent research in blended environments has suggested that online learning environments can result in negative motivational trajectories for under-motivated students (Fryer , Bovee & Nako, 2014). Research has at the same time established that teachers can have a substantial positive effect on this development, even if students' studies are being undertaken mostly outside of class (Fryer & Bovee, 2016, Fryer & Bovee, 2018). Teachers seeking to employ this kind of technology for extra practice during independent study are encouraged to frame the chatbot practice as an opportunity to Learn-more rather than as a convenient tool for practice anywhere and anytime.

For developers of future chatbots specifically for language practice (or at least having that in mind), ensuring that chatbots can adjust (or be adjusted) to students' language competence is essential. Furthermore, consistent with the previous discussion, students need to see chatbots as an opportunity to learn-more (i.e., learn differently) than

they could with a human partner. This is an important aspect of future chatbots that developers might work with educators to enhance.

Users' language competence is a critical issue in human-technology language practice (see Fryer & Bovee, 2018 for an example) that must be addressed at the initial design stage. As a chatbot that is intelligent enough to adapt to learners' levels is still not yet available, the most straightforward approach might simply be to develop a broad range of chatbots (or versions), both for different topics and a range of levels. The second suggestion, that students should perceive chatbots as an opportunity to "Learn-more", means that chatbots need not be a facsimile of human interaction. Some of the reasons students in the current study felt they learned more with the chatbot was precisely because it offered opportunities that human language learning partners couldn't (a broad range of expressions/questions and vocabulary) and/or wouldn't (keep on talking, enable repetitive practice). Students' interest in tasks was the focus of the current study but perceptions of enhanced learning reigned in the mixed-methods analyses. For example, while the reported merit "situational interest" was aligned with above average interest, it was not comparable to the amount of interest when students reported learned-more as the primary merit of the chatbot. This was particularly relevant for the students who reported the primary demerit of the chatbot interaction as communication problems: i.e., 3 SDs difference in chatbot conversation interest. We suggest therefore that ensuring chatbots put helping students learn-more first is a step toward ameliorating communication and technical issues common with chatbots.

9 Limitations and future directions

As with any study conducted within one educational context, the external validity of our results awaits further tests both nationally and internationally. Furthermore, research with other populations such as adults and secondary students are also important. The current study used a mixed-methods design, built on a longitudinal study and utilised a measure of language competency. However, the test was a convenience made possible by the institution's practice of annual standardised examinations for all students. It was also receptive and not productive (i.e., listening and reading, not writing and speaking), which may have limited its predictive power for interest in tasks, which included productive as well as receptive competency. Finally, the study was conducted with a convenience sample of students rather than a random sample from the institution more broadly. Without further experimental research, such as the previous study this research sought to build on (Fryer et al., 2017), the current findings should be treated with caution.

In addition to implementing experimental designs, future studies need to begin developing or adapting current chatbots for the specific purpose of language learning. Chatbots need to be built to address the concerns presented here and in past studies (Coniam, 2014; 2008; Fryer & Nakao, 2009; Fryer & Carpenter, 2006).

10 Conclusions

From the current study we draw a handful of preliminary conclusions. First, that a rebound in interest eventually follows novelty effects, suggesting that a spaced approach to chatbot use might support interest. Second, that for language practice, interest in communication activities like those used in the current study arises chiefly from interest in talking to human partners and interest in the language course broadly. These

wellsprings of interest must be attended to first if educational technology is to play its role in a blended approach to language learning.

The future use and design of chatbots for language practice should take students' competence level into consideration. Teachers might focus on framing the chatbot conversations as an opportunity for students to learn more and different things than one could get from a human language learning partner—rather than the chatbot's convenience. This could mean the scaffolded introduction of new vocabulary, grammar, and expressions, which a human partner is unlikely to present. It could also mean providing consistent understandable repetition, which a human partner is unlikely to want to provide.

In the (not so distant) future, chatbots (or chatbots' great-grandchildren) will fundamentally change how we learn new languages. For the foreseeable future, however, using chatbots will result in a combination of merits and demerits. We suggest that language-learning orientated chatbot design elements and appropriate instructor guided use might forge a balanced path and bring these powerful devices firmly into the language learner's toolbox.

11 References

Allison, D. A. (2013). *The patron-driven library: a practical guide for managing collections and services in the digital age.* Oxford, UK: Chandos Publishing.

Block, N. (1981). Psychologism and behaviorism. In S. Shieber (Ed.), The turing Test: Verbal behavior as the hallmark of intelligence (pp. 229-266). UK: MIT Press.

Bovee, H. N., & Fryer, L. K. (2011). オンラインCALLシステムに関する開発経過報告. 九州産業大学 *COMMON, 31*.

Bradeško, L., & Mladenić, D. (2013). A Survey of Chatbot Systems through a Loebner Prize Competition: Ljubljana Slovenia: Artificial Intelligence laboratory, Jozef Stefan Institute.

Brislin, R. W. 1980. ‘‘Translation and Content Analysis of Oral and Written Materials. Handbook of Cross-cultural Psychology, Vol. 2, edited by H. C. Triandis and J. W. Berry, 389 444. Boston: Allyn & Bacon.

Burns, H.L. & Capps, C.G. (1988). Foundations of Intelligent Tutoring Systems: An Introduction. In M.C. Polson & J.J. Richardson (Eds.), *Foundations of Intelligent Tutoring Systems*, Hillsdale, NJ: Lawrence Erlbaum.

Chen, J. A., Tutwiler, M. S., Metcalf, S. J., Kamarainen, A., Grotzer, T., & Dede, C. (2016). A multi-user virtual environment to support students' self-efficacy and interest in science: A latent growth model analysis. *Learning and Instruction, 41*, 11–22. doi:10.1016/j.learninstruc.2015.09.007

Clark, R. E. (1983). Reconsidering Research on Learning from Media. *Review of Educational Research, 53*, 445-459. doi:http://www.jstor.org/stable/1170217n

Coniam, D. (2008). Evaluating the language resources of chatbots for their potential in English as a second language. *ReCALL, 20*(01), 98-116. doi:doi:10.1017/S0958344008000815

Coniam, D. (2014). The linguistic accuracy of chatbots: usability from an ESL perspective. *Text & Talk, 34*(5), 545-567. doi: 10.1515/text-2014-0018

Dale, R.,(2016). The return of the chatbots. *Natural Language Engineering*, *22*(05), 811–817. http: 10.1017/S1351324916000243

Devellis, R. F. (2012). *Scale Development: Theory and application* (3rd ed.). Thousand Oaks, CA: Sage.

Fryer, L. K., & Ainley, M. (2018). Supporting interest in a study domain: A longitudinal test of the interplay between interest, utility-value, and competence beliefs. *Learning and Instruction*. doi: 10.1016/j.learninstruc.2017.11.002

Fryer, L. K., & Bovee, H. N. (2018). Staying motivated to e-learn: Person- and variable-centred perspectives on the longitudinal risks and support. *Computers & Education, 120*, 227-240. doi: 10.1016/j.compedu.2018.01.006

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. doi: 10.1016/j.chb.2017.05.045

Fryer, L. K., Ainley, M., & Thompson, A. (2016). Modelling the links between students' interest in a domain, the tasks they experience and their interest in a course: Isn't interest what university is all about? *Learning and Individual Differences, 50*, 157-165. doi: 10.1016/j.lindif.2016.08.011

Fryer, L. K., & Bovee, H. N. (2016). Supporting students' motivation for e- learning: Teachers matter on and offline. *Internet and Higher Education*. doi:

10.1016/j.iheduc.2016.03.003

Fryer, L. K. (2015). Predicting self-concept, interest and achievement for first-year students: The seeds of lifelong learning. *Learning and Individual Differences, 38*, 107-144. doi: 10.1016/j.lindif.2015.01.007

Fryer, L. K., Bovee, H. N., & Nakao, K. (2014). E-learning: Reasons students in language learning courses don't want to. *Computers & Education, 74*, 26-36. doi: 10.1016/j.compedu.2014.01.008

Fryer, L. K., Stewart, J., Anderson, C. J., Bovee, H. N., Gibson, A. (2010). Coordinating a vocabulary curriculum: Exploration, pilot, trial and future directions. In A. Stewart (Ed.), *JALT2010 Conference Proceedings*. Tokyo: JALT. Permanent Online Location: http://jalt-publications.org/proceedings/issues/2011-10_2010.1

Fryer, L. K., & Nakao, K. (2009). Asessing chatbots for EFL use. In A. Stoke (Ed.), *JALT2008 Conference Proceedings*. Tokyo: JALT. Permanent Online Location: http://jalt-publications.org/proceedings/articles/84-jalt2009-proceedings-co ntents.

Fryer, L. K., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology, 10*, 8-14. Permanent Online Location: http://llt.msu.edu/vol10num3/emerging/default.html

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. *Information and Systems in Education, 13*, 1-7.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*, 612-618.

Heller, B., Proctor, M., Mah, D., Jewell, L., & Cheung, B. (2005, June). Freudbot: An investigation of chatbot technology in distance education. In *EdMedia: World*

Conference on Educational Media and Technology (pp. 3913-3918). Association for the Advancement of Computing in Education (AACE).

Hidi, S., Renninger, K. A., & Krapp, A. (2004). Interest, a motivational variable that combines affective and cognitive functioning. In *Motivation, emotion, and cognition* (pp. 103-130). Routledge: New York.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111-127. doi:10.1207/s15326985ep41024

Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior, 49*, 245-250. 10.1016/j.chb.2015.02.026

Huang, J., Li, Q., Xue, Y., Cheng, T., Xu, S., Jia, J., & Feng, L. (2015). Teenchat: a chatterbot system for sensing and releasing adolescents' stress. *Health Information Science* (pp. 133-145): Springer.

Huang, P., Lin, X., Lian, Z., Yang, D., Tang, X., Huang, L., . . . Zhang, X. (2014). Ch2R: A Chinese chatter robot for online shopping guide. *CLP 2014*, 26.

Hidi, S., & Ainley, M. (2008). Interest and self-regulation: Relationships between two variables that influence learning. *Motivation and self-regulated learning: Theory, research, and applications*, 77-109

Ichihara, M., & Arai, K. (2004). The development of academic perceived competence and intrinsic interest: A cross-sectional study in Grade 4 through 9 students. Tsukuba Psychological Research, 27, 43–50.

Jia, J., & Chen, W. (2008). Motivate the Learners to Practice English through Playing with Chatbot CSIEC *Technologies for E-Learning and Digital Entertainment* (pp. 180-191). Berlin, Heidelberg: Springer.

Kukulska-Hulme, A., & Shield, L. (2008). An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL, 20*(03), 271-289. doi: 10.1017/S0958344008000335

Lasek, M., & Jessa, S. (2013). Chatbots for customer service on hotels' websites. *Information Systems in Management, 2*, 146-158.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2015). *Introduction to linear regression analysis*: John Wiley & Sons.

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*, 156. doi: 10.1037/0022-0663.94.1.156

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? Cognition and Instruction, 19(2), 177-213. doi: 10.1207/S1532690XCI190202

No Author. (2015). Wikipedia: Cleverbot. Accessed on December 29th, 2016. https://en.wikipedia.org/wiki/Cleverbot

Piaget, J. (1976). Piaget's Theory. In B. Inhelder, H. H. Chipman, & C. Zwingmann (Eds.), *Piaget and His School: A Reader in Developmental Psychology* (pp. 11-23). Berlin, Heidelberg: Springer.

Renninger, K. A., & Hidi, S. (2011). Revisiting the conceptualization, measurement, and generation of interest. *Educational Psychologist, 46*, 168-184. doi:10.1080/00461520.2011.587723

Roussou, M. (2004). Learning by doing and learning through play: an exploration of interactivity in virtual environments for children. *Computers in Entertainment, 2*, 10-10.

SAS. (2007-2011). JMP Version 9.01. Cary, NC: SAS Institute.

Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior, 58*, 278-295. doi:10.1016/j.chb.2016.01.004

Shawar, B. A., & Atwell, E. (2007). *Fostering Language Learner Autonomy Through Adaptive Conversation Tutors.* Paper presented at the Proceedings of the The fourth Corpus Linguistics conference.

Silvia, P. J. (2003). Self-efficacy and interest: Experimental studies of optimal incompetence. *Journal of Vocational Behavior, 62*(2), 237-249. doi:10.1016/S0001-8791(02)00013-1

Stewart, J., Fryer, L. K., & Gibson, A. (2013). Assessing the dimensionality of three hypothesized sub-skills of L2 vocabulary proficiency. *JACET JOURNAL, 51*, 51-71.

Stewart, J., Gibson, A., & Fryer, L. K. (2012). Examining the reliability of a TOEIC Bridge practice test under 1 and 3 parameter item response models. *Shiken Research Bulletin, 16.*

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston:

Pearson Education.

Tobias, S. (1995). Interest and metacognitive word knowledge. *Journal of Educational Psychology, 87*, 399. doi:10.1037/0022-0663.87.3.399

Voogt, J., Fisser, P., & Wright, J. (2015). Computer-assisted instruction. *International Encyclopedia of the Social and Behavioral Sciences.-2nd Ed.-Vol. 4*, 493-497.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the development of children, 23*, 34-41.

Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1). Retrieved from http://i5.nyu.edu/~mm64/x52.9265/january1966.html

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology, 7*, 1182-1189. doi:10.1111/j.1365-2656.2006.01141.x

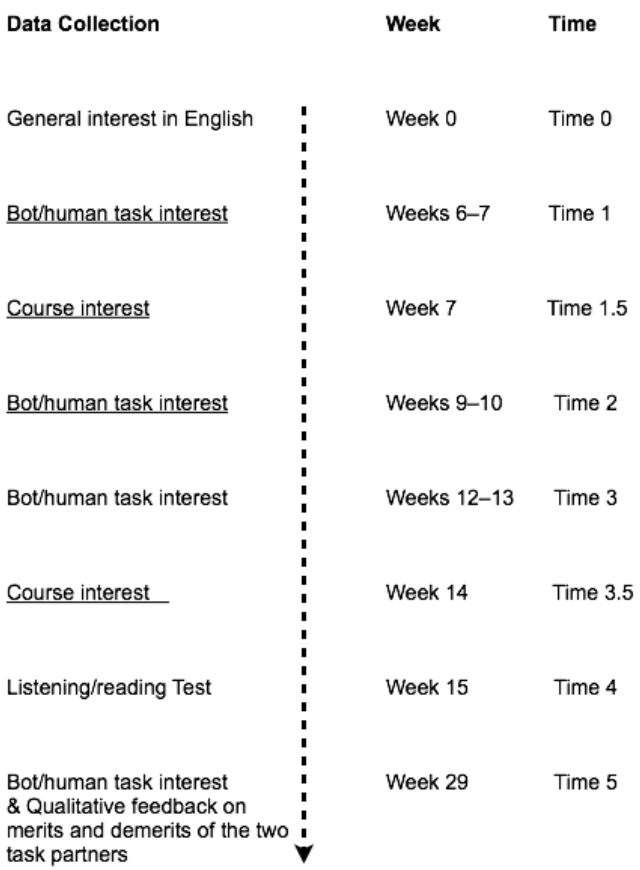

Figure 1. Overall Research Design

Note: Times 1, 1.5, 2 and 3.5 (underlined) were included in the Structural Equation Modelling (Fryer et al., 2017) findings that the current study builds on. For the current study all Time points were drawn upon for the full regression model, Time 4 and 5 for ANOVA tests and Merit/Demerit integrated examination of Chatbot interest and prior competence.

Table 1
Regression results of prior measures of interest and competency for Time 5 human and chatbot partner task interest

	Chatbot conversation interest Time 5	Human conversation interest Time 5
Full model	R^2 = .65, F (10, 81) = 14.90	R^2 = .67, F (10, 91) = 16.02
Final model	R^2 = .61, F (2, 89) = 69.68	R^2 = .66, F (4, 87) = 39.93
Course interest Time 3.5	R^2 = .49, F (1, 90) = 84.97	R^2 = .49, F (1, 90) = 86.95
Chatbot conversation interest Time 1	R^2 = .40, F (1, 90) = 59.63	R^2 = .23, F (1, 90) = 26.82
Human conversation interest Time 1	R^2 = .51, F (1, 90) = 36.54	R^2 = .47, F (1, 90) = 79.08

Note: The final model only includes significant predictors from the full model. Below the final model, the specific contribution to the final model's variance is presented. Significance for all regressions undertaken was $p < .001$.

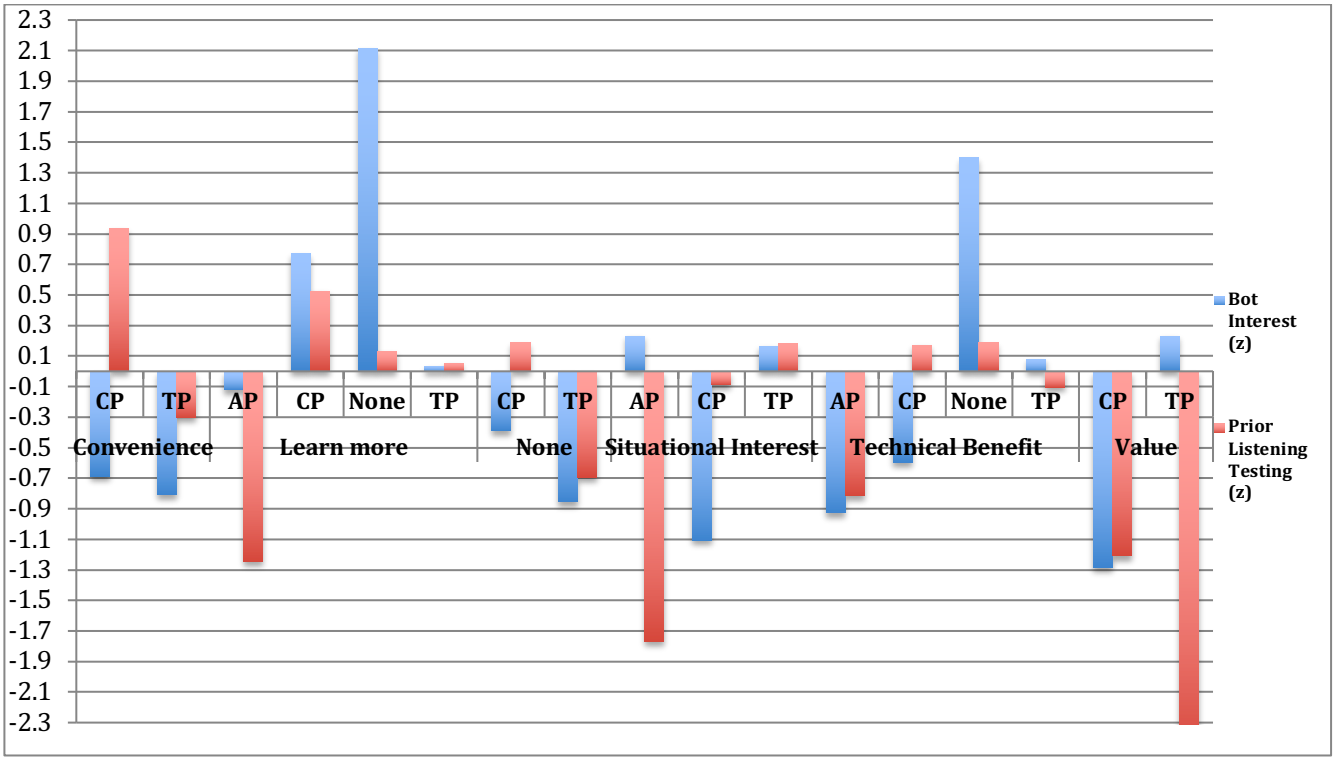

Figure 2. Prior listening competence and interest in chatbot tasks organised by students' reported demerits within their reported merits of chatbots.
Note: Convenience, Learn-more, None, Situational interest, Technical Benefits and Value are students' reported merits for conversations with the chatbot
partners. Organised within these reported merits are the demerits the same students reported: CP = communication problem, TP = Technical problem, AP = ability problem, None = no demerit. All data are presented as Z-scores. For each of these nested demerits within merits, Z-scored interest in the chatbot speaking task and Z-scored prior language exam results are presented.

12 Appendixes

Table 1.

Correlations, descriptive statistics and Cronbach's Alphas

	Human T1	Human T2	Human T3	Chatbot T1	Chatbot T2	Chatbot T3	Domain T0	CourseT 1.5	CourseT 3.5	Test T4	Human T5	Chatbot T5
HumanT1												
HumanT2	.73**											
HumanT3	.70**	.80**										
ChatbotT1	.74**	.67**	.63**									
ChatbotT2	.66**	.83**	.80**	.62**								
ChatbotT3	.69**	.68*	.71**	.69*	.71**							
DomainT0	.39**	.43**	.44**	.30**	.51**	.35**						
CourseT1.5	.66**	.54**	.57**	.56**	.56**	.59**	.43**					
CourseT3.5	.48**	.48**	.46**	.44**	.45**	.52**	.46**	.61**				
TestT4	.15	.18	.21*	.01	.29**	.21*	.26*	.23*	.29**			
HumanT5	.68**	.65**	.67**	.49**	.62**	.54**	.44*	.51**	.73**	.18*		
ChatbotT5	.64**	.66**	.72**	.54**	.65**	.69**	.35**	.53**	.71**	.19*	.75**	
Mean	3.87	3.78	3.72	3.80	3.47	3.38	3.90	4.07	3.88	16.42	3.89	3.64
SD	.96	1.05	1.07	1.10	1.17	1.26	0.90	1.13	1.00	3.78	.99	1.12
Cronbach's Alpha	.93	.95	.95	.96	.96	.95	.85	.93	.93	.90	.94	.94

Note: T1= "Time 1, T1.5= Time 1.5, T2= Time 2, T3= Time 3, T3.5= Time 3.5, T4= "Time 4, T5= Time 5. Human refers to human-partnered task interest. Chatbot refers to chatbot-partnered task interest. Course refers to course interest. Domain refers to domain-level interest (i.e., interest in learning English as language generally). Test refers to a standardised achievement test. The test results range from 0 to 20. * = $p < .01$, ** = $p < .001$.

Table 2. Counts, means and SD for each Human Merit/Demerit and ANOVA results across the coded categories

	Chatbot interest	SD	Human interest	SD	Prior listening test	SD	N	R2 Interest (F, p)	R2 Fluency (F, p)
Human conversation merits								.11 (F= 3.13, *p*=.018)	.03(F= .08, *p*=.49)
Communication ease			3.62ab	0.96	77.71a	16.77	42		
Learn-More			4.29a	1.20	79.95a	15.90	19		
None			2.94a	1.16	70.00a	29.59	9		
Social Benefits			3.87ab	0.77	76.32a	19.79	22		
Situational Interest			4.13ab	0.97	70.56a	18.22	19		
Human conversation demerits								.02 (F= .31, *p*=.92)	.02(F= .04, *p*=.87)
Ability problems			3.60a	0.53	72.67a	14.29	3		
Communication problems			3.76a	1.27	75.93a	13.69	17		
Not interesting			3.60a	0.57	68.50a	14.85	2		
No good for learning			3.92a	1.14	82.50a	13.74	13		
None			3.87a	1.02	74.24a	21.21	40		
No Value			3.20a	2.55	70.50a	3.54	3		
Social demerits			3.95a	0.78	77.03a	20.02	33		
Chatbot conversation Merits								.09 (F= 1.80, *p*=.11)	.12 (F=2.70, *p*=.025)
Convenience	2.78a	1.12			79.80a	17.29	10		
Learn-More	3.86a	1.05			77.27a	19.13	22		
None	2.77a	1.39			68.83b	10.17	7		
Situational interest	3.77a	1.26			78.52a	14.02	25		
Technical benefits	3.55a	1.12			76.75a	19.61	44		
Useful	3.33a	0.99			41.33b	19.86	3		
Chatbot conversation demerits								.09 (F= 3.27, *p*=.024)	.11 (F=4.04, *p*=.009)
Ability problems	3.23ab	0.71			56.88b	34.26	8		
Communication problems	3.08b	1.24			83.33a	11.81	15		
None	5.60a	0.57			81.50ab	0.71	2		
Technical problems	3.61ab	1.15			76.49a	16.52	86		

Note: Means within merits and demerits, tested for human and then chatbot partners, are significantly different ($p < .05$) where the letter nomenclatures (a, b, ab) are different: i.e., if two means have the same letter nomenclatures they are not significantly different ($p < .05$).