Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners

Luke K. Fryer, The University of Hong Kong; The University of Sydney
lukefryer@yahoo.com (**corresponding author**)

Mary Ainley, University of Melbourne
maryda@unimelb.edu.au

Andrew Thompson, Kyushu Sangyo University, Language Education and Research Centre
thompson@ip.kyusan-u.ac.jp

Aaron Gibson, Kyushu Sangyo University, Language Education and Research Centre
aaronlgibson@gmail.com

Zelinda Sherlock, Kyushu Sangyo University, Language Education and Research Centre,
sherlock@ip.kyusan-u.ac.jp

Funding:
The first author's contribution was partially funded by a Thomas and Mary Ethel Ewing Scholarship

Acknowledgements:
We would like to acknowledge the financial support of Kyushu Sangyo University's Computer Network Centre.

FULL REFERENCE:

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:https://doi.org/10.1016/j.chb.2017.05.045

This is the accepted version prior to journal formatting.

The official version can found at:

https://authors.elsevier.com/a/1V8wS2f~UW0xXp**-->FREE COPY the end of JULY 2017.**

Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners

Abstract

Novel technology can be a powerful tool for enhancing students' interest in many learning domains. However, the sustainability and overall impact of such interest is unclear. This study tests the longer-term effects of technology on students' task and course interest. The experimental study was conducted with students in foreign language classes (n=122): a 12-week experimental trial that included pre- and post-course interest, and a sequence of task interest measures. Employing a counterbalanced design, at three week intervals students engaged in separate speaking tasks with each of a Human and "Chatbot" partner. Students' interest in successive tasks and in the course (pre-post), were used to assess differential partner effects and course interest development trajectories. Comparisons of task interest under different partner conditions over time indicated a significant drop in students' task interest with the Chatbot but not Human partner. After accounting for initial course interest, Structural Equation Modelling indicated that only task interest with the Human partner contributed to developing course interest. While Human partner task interest predicted future course interest, task interest under Chatbot partner conditions did not. Under Chatbot partner conditions there was a drop in task interest after the first task: a novelty effect. Implications for theory and practice are discussed.

1 Introduction

At the heart of becoming competent in any domain stands the necessity for persistence. While there is a broad range of theories modelling how such persistence is achieved, developing interest in the domain is one approach, which is supported by both research (Ainley, Hidi, & Berndorff, 2002; Tobias, 1995) and common-sense. As a result, supporting and, where necessary, stimulating students' interest is an implicit part of every educator's belief.

The question is how interest might be stimulated most effectively. Two approaches that have received considerable attention are the role of perceived value (e.g., Hulleman, Godes, Hendricks & Harackiewicz, 2010) and of curriculum tasks (Hanus & Fox, 2015; Guberman, & Leikin, 2012). In the context of foreign language learning, there is a longstanding focus on the importance of creating tasks that support sustained learning (e.g., Lightbown & Spada, 1994). Recently, research attention both in the area of language learning and general education, has focused on the potential of technological tools to enhance classroom motivation and thereby learning. One technology that has been suggested as a potentially powerful tool for enhancing students' language learning efforts is the area of Chatbots (Goda, et al., 2014; Stickler & Hampel, 2015; Fryer, 2006; Fryer & Nakao 2008; Conaim, 2008). Chatbots are software avatars with limited, but growing capability for conversation with human beings.

However, in the context of technology-based educational interventions, current research (e.g., Chen et al., 2016) has raised concerns regarding the potential for novelty effects to mask the real impact of technological interventions. As a result, the only confident means of assessing the potential of Chatbots as a tool for enhancing interest in

language learning courses is an experimental trial. In the current research an experimental trial was conducted to compare the influence of Chatbot and Human partners on both task interest and later course interest. This study was undertaken within the context of a university language course using a framework that distinguishes interest for task, for course and for domain (Fryer, Ainley & Thompson, 2016) when modelling interest development.

1.1 Interest development

From its transition across philosophy to psychology, to its strong empirical impact on reading research, our understanding of interest as a psychological construct has a considerable history (see e.g., Hidi, 1990). It has long been recognised that there are at least two different types of interest; situational and individual. The labelling of these types has varied over time and between researchers. However, these two types have generally been identified as an early stage or phase which is transitory and chiefly affective. This early stage, sometimes separated into an emerging situational interest and a stabilized situational interest (Krapp & Prenzel, 2011), is then potentially followed by a stage that is longer-lasting, and includes additional value and epistemological components (Schiefele, 1991).

One widely-cited framework for understanding the development of interest is the Four-Phase Model of Interest Development (Hidi & Renninger, 2006; Renninger & Hidi, 2011). This model describes the potential development of an individual's interest from initially stimulated interest in a topic - triggered situational interest. If interest is sustained, and allowed to grow, then triggered situational interest develops into the second phase of

maintained situational interest. The later two phases of development in this model are described as emerging individual interest and well-developed individual interest.

1.1.1 Related educational principles

The Four-Phase Model of Interest Development suggests to educators a broad path that learners might travel from initial triggering of interest to a sustainable personal interest in a domain of study. Hidi and Renninger (2006) emphasise that the length of each phase is variable and that an individual's interest development might cease at anytime. The instructional environment plays a role in triggering situational interest through a range of novel and social activities. The maintenance and deepening of interest across the remaining three phases consists chiefly of supporting personal involvement, knowledge development and increasing value of the domain.

1.2 Interest development in formal education

When the focus is on understanding the development of interest in domains across specific university courses, a model of interest development that distinguishes three levels has been suggested (Fryer, Ainley & Thompson, 2016). The first level relates to the specific tasks which represent learning events such as lectures, group projects, independent reading, watching videos, and doing experiments. The second level relates to students' interest in the course itself. The final level is their interest in the broader study domain. Some initial research using this framework on interest development reported that course interest mediated the relationship between students' interest in tasks and their interest in the broader study domain. This result makes stimulating and sustaining course interest of substantial importance if university instructors are seeking to encourage students to continue with further studies in the domain. Essentially, these results suggest

that tasks matter because they directly build interest in courses which can directly impact interest in study domains. Hence, further research into task features that stimulate and sustain interest is warranted.

In the current environment where much of the educational innovation is technology orientated, an important direction for research is to assess the potential for technological learning tools to enhance students' interest in curriculum tasks.

1.3 Technology for enhancing interest and learning

The growing use of technology has long been heralded as a means to dramatically shift our understanding of education, however not always in the ways we might expect (Naisbitt & Cracknell, 1984). Futurists, just a few decades ago, pointed to skills we would need in a world filled with omnipotent computers, while others underlined the importance of the growing constructivist movement for meaningful learning in any age (Nickerson, 1988). Few trends in educational technology have been more closely watched than the steady growth of intelligent tutors within the field of artificial intelligence (AI). In the broad array of roles intelligent tutors are able to perform, they are at the cutting-edge of human-technology interaction. Arising out of Computer Assisted Instruction (CAI), early attempts at intelligent tutors (e.g., Carbonell, 1970) initially aimed to anticipate rather than interact with learners. Since the time of the initial attempts at CAI, many educational researchers have collaborated with technologists in the relentless pursuit of smart education. From virtual tutors and coaches to virtual environments and the broad appeal of game based learning, intelligent tutors seem here to stay. Early studies (e.g., Lester et al., 1997) pointed to the positive effect that basic "life-like" agents could have on learners' perceptions of learning environments. Steady

progress in the design of these educational agents coupled with research into their effectiveness has both provided support for their broad motivational benefits and refined our understanding of how they support learning. Keystone research in this field by Mayer and colleagues (e.g., Mayer, Dow, & Mayer, 2003; Moreno, Mayer, Spires, & Lester, 2001) has demonstrated that for university students working with physics problems, the intelligent agent was more effective when explanations to the student were in the form of speech rather than on-screen text. Furthermore, this research by Mayer and colleagues found that visual representations of the intelligent tutor did not significantly support increased learning outcomes. More recent phenomenological (Veletsianos & Miller, 2008) and experimental (Veletsianos, 2010) research examining conversational and pedagogical agents have posed a more nuanced set of questions regarding visual interaction between digital agent and human participant. These questions now go beyond considering intelligent tutors as instructive tools, to questions of how humans might interact and carryout meaningful communication with the intelligent agents.

From an educative perspective, the step from agents that support learning to agents that communicate with humans opens up possibilities in the area of language learning. In few areas of education have the advances of technology been more acutely felt than second and foreign language-learning (Blake, 2013). While the audio/visual support that technology provides is important for all education, the possibility of conversational interaction with an intelligent agent is at the heart of technology's potential contribution to language learning. It is widely acknowledged that massive amounts of comprehensible language input and practice are essential for meaningful language learning to take place. Across Asia, for example, the low number of native

speakers of English puts a premium on opportunities for students to practice when learning this new language. One technological response to this problem is the potential of "Chatbots" or intelligent agents for conversational practice, which are online software capable of carrying on a conversation with interest humans.

Consistent with much of the intelligent tutor research (Johnson & Lester, 2016), students have reported motivational benefits with Chatbots during a classroom task (Fryer, 2006). This early text-based study suggested that many students were more comfortable trying a new language with a Chatbot than with a Human partner. However, further research with Chatbots (Coniam, 2008; Fryer & Nakao, 2008), pointed to the text-only nature of this interaction as a factor restricting the usefulness of Chatbots for many language learners. Along with questions of usefulness, these authors highlighted the inauthentic nature of text-based Chatbots as a source of conversational practice language students. Despite these issues, both early (Weizenbaum, 1966) and very recent (Hasler, Tuchman, & Friedman, 2013; Hill, Ford, & Farreras, 2015) research with text-based Human-Chatbot interactions have consistently pointed to their potential benefits, particularly with regard to the motivation they seem to inspire in their users. Furthermore, research seeking to build directly on Weizenbaum' original efforts has suggested that current Chatbots have and continue to improve as many of them learn from their "round-the-clock" web-based interaction with interested humans from around the globe (Shah, Warwick, Vallverdú, & Wu, 2016).

Recent advances by Chatbot developers and text-to-speech/speech-to-text software have begun to make spoken Human-Chatbot interaction a growing option, opening up new possibilities for Human-Chatbot interaction and learning. Through a

broad range of devices it is now possible to talk directly to Chatbots, who in turn are able to respond in an engaging manner via text or audio. The potential role of these Chatbots for dramatically expanding students' opportunities for language interaction is as yet untapped. In addition to their language practice opportunities, Chatbots might also have a role to play in triggering students' interest in language learning thereby contributing to sustaining interest in learning the new language beyond a single course of study.

However, before drawing conclusions regarding the effectiveness of Chatbots as interactive partners in language learning, the issue of potential novelty effects confounding assessment of increases in motivation needs to be examined. Novelty effects occur when the simple newness of a technology causes a rise in motivation or in achievement. Novelty effects have a considerable history within educational technology research and have been acknowledged as far back as the 1960s (see Clark, 1983) and also figure in recent publications (see e.g., Chen et al., 2016). To avoid simple novelty effects, research needs to be designed to extend over sufficient time to allow potential novelty effects to diminish and the enduring effect of the technology to be assessed (Bracht & Glass, 1968). Hence, an experimental trial that continues over an extended period of time is one approach to assess Chatbot usefulness after any novelty effects have diminished. In this way the advantages of technological innovations for the learning of foreign languages can be more accurately determined.

1.4 Current Study

Across the globe, the number of students learning English within formal educational institutions is dramatically increasing. In Japan, English education starts in elementary school and is compulsory right through to university. However, due primarily to its

limited usefulness with students' future work and private lives, and the relatively low value of English in Japan, (Fryer, Carter, Ozono & Anderson, 2013; Matsuda, 2009), students can be forgiven for sometimes not seeing the clear and present value of learning the language. This lack of perceived value and the fact that English language courses are compulsory in Japanese universities, contribute to the low levels of interest students have for both their English courses and for the general English language domain (Fryer, 2015).

Concerns regarding student interest in learning the English language and the necessity of expanding opportunities for language use converge to provide the rationale for testing the potential role of Chatbot partners in English language learning. Hence an experimental test of the effectiveness of Chatbot partners for increasing interest in language learning was conducted across a twelve-week language course with a control condition of Human partners for the same tasks. We were interested in both the overall level of students' interest at the task level and the longer-term implications for interest in the language course. For the present test we utilised Cleverbot (Carpenter, n.d.). Cleverbot is software designed to learn from its conversations with humans—more than 200 million to date (Wikipedia, n.d.). It draws on past interactions to determine future questions and answers. Based on previous studies, Cleverbot is useful for motivating foreign language students (Fryer, 2006), as well as general users (Hill et al. 2015), to communicate. Analysis of Cleverbot's interactions (Conaim, 2008) has also demonstrated that it is a grammatically clear conversationalist. Furthermore, the Chatbot Cleverbot was based on (Jabberwacky), which won the Loebner contest twice. The Loebner contest is an annual competition testing whether competing Chatbots' responses are indistinguishable from human responses. The Chatbot closest to this goal each year wins a bronze medal.

2 Aims

The current experimental study examined the difference between students' interest in classroom speaking tasks under conditions of Chatbot and Human partners. Using a counterbalanced partner design (Group 1: Chatbot/Human, Group 2: Human/Chatbot) the same speaking task was repeated with the different partner after one week. At three-week intervals this procedure was repeated with a two further speaking tasks across the twelve week English as a foreign language course. In addition, this study investigated the longer-term implications of task interest with different partners (Chatbot and Human) for students' interest in the broader language course.

Based on past research (Fryer, 2006), we predicted that initially students would be more interested in spoken language learning tasks with an unfamiliar Chatbot partner than with a Human partner. Hence, to reduce the novelty of the Chatbot we gave all participating students a familiarizing experience with the Chatbot prior to the study. If Chatbots are an effective partner for the specific language learning task, students' task interest will be sustained across the task under the Chatbot condition.

With regard to the longer-term effects of interaction with Chatbot and Human partners on students' interest in the course, we predicted that interest in speaking tasks conducted with both Chatbot and Human partners would make a positive contribution to interest in the course.

3 Methods

3.1 Sample and context

The current study was undertaken within first- and second-year compulsory English as a foreign language classes at one private university in Japan. Students (n = 122) from five faculties participated in the study. Participating students attended two classes a week. Students' classes were embedded within a coordinated program of study. Consistent classroom materials, weekly e-learning assignments and assessment were employed across all participating classes.

3.2 Instrumentation

Two Likert-type scales were used in the current study, course interest and speaking task interest. All items required ratings from 1 = "nothing like me" to 6 = "totally like me". The course interest scale consisted of four items, for example, "I am fully focused on learning English in this course" and "This English course is interesting". The speaking task interest scale consisted of five items, for example, "This activity is personally meaningful" and "I enjoyed learning English in this activity". These measures have been used in a previous study (Fryer, Ainley & Thompson, 2016) where strong convergent and divergent validity, and reliability (>.7; Devellis, 2012) were reported. See Table 1 for Cronbach's Alpha results from the current study.

3.3 Research Design

In the current study each participating class (six in total) was randomly divided into two groups. Three weeks prior to the commencement of the study (T0) students were introduced to the Chatbot technology. Three weeks later (T1) half of the class (Group 1)

completed a prepared speaking task (Task 1, part A) with a Human partner and half of the class (Group 2) used a tablet to complete the same speaking task with a "Chatbot" partner. At the end of the task students were asked to report their interest in the task they had just completed. The following week, the treatments were reversed (Task 1, part B), and the same speaking task and task interest procedures repeated. The first course interest scale was administered one week later.

==========Figure 1 ABOUT HERE==========

Three weeks after the initial speaking task (T2) the procedure was repeated with a new speaking task (Task 2, part A and one week later part B). This was repeated again three weeks later (T3) when students completed Task 3, part A, and a week later part B. The final course interest scale was repeated immediately after the final speaking task. Figure 1 summarizes this research design.

3.4 Analyses

All latent analyses (utilising measurement models based on the scale items, not mean scores or sum scores of scales) were undertaken with M*plus* 7.0 (Muthén & Muthén, 1998-2013) and analyses with observed (difference test of scale means) variables were conducted with JMP 9.01 (SAS, 2007-2011). Analyses began by testing for order effects comparing T1 Chatbot scores for Group 1 with Chatbot scores for Group 2 (Chatbot administered first vs. second). In the same way any order effect for the T1 Human scores was established. If there were no significant order effects, scores for the Group 1 and

Group 2 Chatbot conditions and the Group 1 and Group 2 Human conditions could be combined in all further analyses. Reliability of the scales and an examination of the correlations and descriptive statistics between task interest and course interest measures were then conducted. Differences between task interest scores under Chatbot and Human partner conditions at the three time points were then assessed. Analyses concluded with structural equation modelling to assess the longitudinal relationship between task interest under Human and Chatbot conditions and their predictive relation with later course interest.

In consideration of the relatively small sample size, latent modelling with just two of the three task interest data points was pursued rather than a path analysis of all data points using observed variables. Path analysis would result in the examination of a saturated model and therefore prevent the use of model fit statistics. Furthermore, latent (rather than mean-based observed) measurement has been suggested as important for cross-lagged analyses such as those proposed here (Pedhazur & Pedhazur, 1991). As a result, limited fit indices for a smaller (latent) model were preferred over a larger more complex model with no direction regarding the fit of the tested model to the data. Building on Feinian et al. (2008), Kenny, Kaniskan, and McCoach (2015) have demonstrated that RMSEA is not a useful fit statistic for small sample SEM analyses. As a result analyses in the current study relied on the Comparative Fit Index (CFI) and Standardized Root Mean Square Residual (SRMR). For CFI >.90 and >.95 were held to represent acceptable and good fit (McDonald & Marsh, 1990). For SRMR < .08 represents good fit (Hu & Bentler, 1999).

4 Results

4.1 Test for order effects

ANOVA was used to test for order effects, that is, any effect of Chatbot or Human partner being the first (A) of the two tests in the counterbalance design (see Figure 1). There was no significant difference in task interest scores at T1 based on whether the Chatbot (Group 1, A: $M = 3.67$, Group 2, B: $M = 3.88$, $F(1,120) = 1.03$, $p>.05$) or Human (Group 1, B: $M = 3.85$, Group 2, A $M = 3.91$, $F(1,120) = .1035$, $p>.05$) partner condition was administered first. Similarly, there were no significant order effects at T2 and T3. As a result, A and B speaking task interest scores for each of the Chatbot and Human partner conditions for Time-1, Time-2 and Time-3 were combined for further difference and predictive testing.

4.2 Descriptive findings

The latent correlations, descriptive statistics and Cronbach's alphas for each of the task interest and course interest measures are summarized in Table 1. The strong positive correlations are consistent with the nature of the constructs and their relative temporal distance. All of the mean scores except for the Chatbot condition at Time 2 and Time 3 were above the midpoint (3.5) of the range indicating that most students reported being interested in the tasks. The reliability of all scales was well above what is generally considered to be acceptable (i.e., > .70; Devellis, 2012; see Table 1).

==========Table 1 ABOUT HERE==========

4.3 Differences in task interest: Chatbot versus Human partner

Testing for differences in task interest over the three time points (T1, T2 and T3) for the Human and Chatbot partner conditions proceeded with 3x2 Mixed Design ANOVA test. All of the component means are presented in Table 2. There was a significant difference across the three time points ($F(2,241) = 17.443$, $p<.05$), between Chatbot and Human partner conditions ($F(1,241) = 4.034$, $p<.05$), and an interaction effect between time and partner condition ($F(1,241) = 6.02$, $p<.05$). As can be seen from Table 2, the main effect for time was a significant difference between T1 and both T2 and T3. Pairwise tests (Tukeys HSD, $p < .05$ with Bonferonni adjustment) were used to identify the direction of the interaction effect. There was no significant difference at T1 between task interest mean scores for Human and Chabot partner conditions, however the mean task interest scores for the Human partner condition were significantly higher than task interest for the Chatbot partner condition at both T2 and T3. No significant differences were observed across the three task interest mean scores for the Human partner condition. The mean task interest score for the Chabot partner condition at T1 was significantly higher than task interest for the Chatbot partner conditions at both T2 and T3, which were not significantly different from each other.

==========Table 2 ABOUT HERE ==========

4.4 Model test

To test for the contribution of task interest to course interest over time, a Structural Equation Model was constructed. The model to be tested included only two of the three speaking task interest scores. Given that there was no significant difference between the

scores recorded at T2 and T3 (see Table 1), and that the T3, part B speaking task scores were recorded at the same sitting as the final course interest measure, T2 scores were used to construct the longitudinal model. Given the small sample size, fit for the model was acceptable based on three statistics: CFI = .91, SRMR = .061, Chi-square test = 668.05 (DF = 320, $p < .001$). RMSEA (= .095) was high as expected, but ruled out as an effective fit statistic for the current study based on past evaluations of its performance with smaller sample sizes (Kenny et al., 2015). The full model is presented in Figure 2. All tested predictive relationships were significant ($p < .01$) except for the relationship between task interest with the Chatbot partner at T2 and course interest at T3.

==========Figure 2 ABOUT HERE==========

As Figure 2 shows, there were large significant auto and cross-lagged predictive effects between the Chatbot and Human conditions from T1 to T2. From the T2 partner conditions, only task interest under the Human partner condition significantly predicted T3 course interest. As expected, T1 course interest predicted future course interest (T3).

5 Discussion

The present study was a longitudinal experimental comparison of two speaking tasks in the context of a compulsory English as a foreign language course at a Japanese university. A 3x2 Mixed Design test of interest in the speaking tasks with both Human and Chatbot partners indicated that there was a significant decline in task interest for the Chatbot partner condition. This decline occurred between the first and the second tasks suggesting a novelty effect when interacting with the Chatbot partner. This apparent novelty effect did not occur when interacting with a human partner. A model test of the predictive paths across the study demonstrated that after accounting for prior course

interest, task interest stimulated and sustained when interacting with a Human partner but not with a Chatbot partner significantly contributed to later interest in the course.

5.1 Implications for theory

Two key implications for theory arise from the current study with students learning a foreign language. First, despite providing students with an opportunity to play with the Chatbot prior to the experimental trial, students' interest in the Chatbot partner task significantly decreased. At the same time, interest in the task interacting with the Human partner remained consistently high across all three tasks. While qualitative research is necessary to understand the drop in task interest for the Chatbot partner group, it seems safe to suggest two possible reasons at this stage. The first reason could be a simple novelty effect of the type described by Chen et al. (2016) in their technology-centred intervention. The second is the possibility that authenticity played a role. After one task interaction with the Chatbot partner, students may have perceived this as an inauthentic speaking experience. As a consequence they may have interpreted interaction with the Chatbot partner as a poorer learning experience. Given the fact that all students also had experience with the Human partner condition it is highly likely that some form of comparative evaluation of the two conditions has occurred and the Chatbot partner has been evaluated as a poorer learning partner. The second implication for theory is related to the model of interest operating at task, course and domain levels when considering its contribution in formal educational contexts (Fryer, Ainley & Thompson, 2016). In the current study our longitudinal modelling has demonstrated that in addition to a potential novelty effect, the level of task interest stimulated and sustained under the Chatbot partner condition did not predict future interest in the course. Our previous results suggest

that this is also likely to be the case when considering development of interest in the broader domain.

The current findings do not support the longstanding (Weizenbaum, 1966; Fryer, 2006) and recent (Hill, Ford, & Farreras, 2015) assumptions regarding the motivational benefits of Chatbot interaction. Those assumptions, however, were based on studies examining text exchange interactions with Chatbot partners. It seems relevant, therefore, to point to the importance of the specific task for understanding the usefulness of technology. In the context of the current study, research had suggested that Chatbots were a potential source of motivation for sustained communication to use a foreign language. However, by implementing a longitudinal experimental design with a Human partner control, it appears that past results with Chatbots might not necessarily generalize to oral communication. These results also point beyond the current test with student using a foreign language to suggest that some tasks despite eliciting considerable behavioural interest initially, might not sustain sufficient interest to impact later interest in the broader domain of study.

5.2 Implications for practice

The use of technology within classrooms at all levels expands as costs plummet and these tools become easier to use. The upcoming generation of teachers are digital natives and as a result might be less questioning of technology in the classroom (Lei, 2009). At the same time, there is growing concern about student motivation within formal education and how educators might support 21st century students (Hidi & Harackiewicz, 2000). In the current technological and motivational climate, the growing number of digital native instructors might be inclined to see technology as the answer to stimulating

and sustaining students' interest. The current results suggest the character of the task is critically important for how students interact with tasks delivered through technological interventions. Instuctors and teachers at all levels should be aware that interest and its longterm development go beyond what they can see behaviorally. There are deeper connections which relate to whether students value learning tasks. In addition, the match between the technological innovation and the task requirements is likely to play a role as has been demonstrated in this study focusing on a language learning context. Further research, in particular experimental comparisons of learning conditions are necessary to provide educators with the information they need to make decisions about how and when to effectively use technology to stimulate and then support the development of interest in the course of study and also the broader study domain.

For Chatbot developers seeking to overcome the potential novelty-effects presented by the current study, it is significant first to remember that novelty is an important (initial) component of interest development. For Chatbot interactions to result in sustained interest and substantive learning, however, they need to go beyond novelty and on toward enduring interest development as suggested by the four-phase model of interest development (Hidi & Renninger, 2006; Renninger & Hidi, 2011). Furthermore, research examining the role of novelty within memory suggests that familiarity (rather than novelty) plays an important role in remembering materials (e.g., Poppenk, Köhler & Moscovitch, 2010).

A Chatbot which learners logged into and therefore remembered the users' past questions and level of language use, could, over a series of interactions, become familiar to users. The Chatbot could reuse past language that has been successfully responded to,

thereby enhancing users' self-perceived ability—itself an important support for interest development. The Chatbot could also be programmed to find out what the user was interested in and focus on these topics. Finally, it could remind the user of the necessity of practice and trying new words/phrases, while stressing the importance of making mistakes along the way. Both of these final suggested additions could support students' value for future interactions with the Chatbot—value is also essential for interest development. In sum, current and future Chatbot developers, by attending to our growing knowledge about how humans get interested, could take significant steps toward ensuring that novelty, while certainly a part of initial interactions with Chatbots, does not define the user's final experience.

6 Limitations and Future directions

Consistent with all studies carried out in one specific context, the external validity of our results can only be verified after replication at other institutions, levels of education and in domains other than foreign language learning. Furthermore, research in other cultures is also called for to ensure that these findings do not represent something specific to Japan. Despite its experimental nature, it is not possible to control for all possible influences and it is important to point out that the predictive effects identified in this study are not the same as causation.

It is possible that the counter-balanced design, although a robust means of obtaining within-student comparisons, could have influenced the study's results. Working with Human and then Chatbot partners (or the other way around) might have increased the chances of students directly comparing the two and therefore led, in part, to students' declining interest in the Chatbot. To ensure that this was not the case, future studies

might use a more straightforward experimental design and then compare between students, while still controlling for initial interest. We also suggest that the current study be followed up with qualitative research seeking to understand how students perceived the task when delivered under the different partner conditions, both on first engagement with the task and after repeated trials.

7 Conclusions

The present experimental trial of interest stimulated and sustained with a speaking task delivered under Chatbot and Human partner conditions has two main conclusions. First, novelty effects appear to be a significant issue with technology enhanced tasks like the one employed in this study. Second, tasks seeking to stimulate task interest, and apparently succeeding, might be no more than novelty effects and therefore be unlikely to contribute to students' broader, more long-term interest in the domain. Educators need to carefully reflect on these issues when considering the use of the new tools that technology develops at an ever accelerated pace.

References

Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal Educational Psychology, 94*, 545-561. doi:10.1037//0022-0663.94.3.545

Blake, R. J. (2013). *Brave new digital classroom: Technology and foreign language learning*. Georgetown: Georgetown University Press.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5*, 437-474.

Carpenter, R. (n.d.). www.cleverbot.com. Retrevied on January 22, 2017.

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

Carbonell, J. R. (1970). AI in CAI: An artificial-intelligence approach to computer-assisted instruction. *IEEE transactions on man-machine systems, 11*, 190-202.

Clark, R. E. (1983). Reconsidering Research on Learning from Media. *Review of Educational Research, 53*(4), 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

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

Feinian C., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models. *Sociological Methods & Research, 36*(4), 462-494. doi:10.1177/0049124108314720

Fryer, L. K., Ginns, P. & Walker, R. A. (2016). Reciprocal modelling of students' regulation strategies and motivational deficits for studying. *Learning and Individual Differences*. 51, 220-228. doi: 10.1016/j.lindif.2016.08.032

Fryer, L. K. (2015). Predicting self-concept, interest and achievement for first-year university students: The seeds of lifelong learning. *Learning and Individual Differences*, 38, 107-114. doi: 10.1016/j.lindif.2015.01.007

Fryer. L. K., Carter, P., Ozono, S., Nakao, K., & Anderson, C. J. (2013) Instrumental reasons for studying in compulsory English courses: I didn't come to university to study English, so why should I? *Innovation in Language Learning and Teaching*. 8, 239-256 doi: 10.1080/17501229.2013.835314

Fryer, L. K., & Nakao, K. (2009). Online English practice for Japanese University students: Assessing chatbots. Paper presented at The Japan Association for Language Teaching national conference, Tokyo. Permanent Online Location: http://jalt-publications.org/proceedings/articles/84-jalt2009-proceedings-contents

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

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. doi:10.12937/ejsise.13.1

Guberman, R., & Leikin, R. (2012). *Interesting and difficult mathematical problems: changing teachers' views by employing multiple-solution tasks* (Vol. 16).

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education, 80*, 152-161. doi:10.1016/j.compedu.2014.08.019

Hasler, B. S., Tuchman, P., & Friedman, D. (2013). Virtual research assistants: Replacing human interviewers by automated avatars in virtual worlds. Computers in Human Behavior, 29, 1608–1616.

Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research, 60*, 549–571. doi:10.3102/00346543060004549

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70*, 151-179.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111-127. doi:10.1207/s15326985ep4102_4

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.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling, 6*, 1-55. doi:dx.doi.org/10.1080/10705519909540118

Hulleman, C. S., Godes, O., Hendricks, B., & Harackiewicz, J. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology, 102*, 880.

Johnson, W. L., & Lester, J. C. (2016). Face-to-Face Interaction with Pedagogical Agents, Twenty Years Later. *International Journal of Artificial Intelligence in Education, 26*, 25-36.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The Performance of RMSEA in Models With Small Degrees of Freedom. *Sociological Methods & Research, 44* 486–507.

Lei, J. (2009). Digital Natives As Preservice Teachers. *Journal of Computing in Teacher Education, 25*(3), 87-97. doi:10.1080/10402454.2009.10784615

Lightbown, P. M., & Spada, N. (1994). An Innovative Program for Primary ESL Students in Quebec. *TESOL Quarterly, 28*(3), 563-579. doi:10.2307/3587308

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., B.A., S., & Bhogal, R. S. (1997). *The persona effect: Affective impact of animated pedagogical agents.* . Paper presented at the CHI'97.

Matsuda, A. (2009). Desirable but not necessary? The place of World Englishes and English as an international language in English teacher preparation programs in Japan. *English as an international language: Perspectives and pedagogical issues*, F. Sharifian. 169-189. New York: Multilingual matters.

Mayer, R. E., Dow, G. T., & Mayer, S. (2003). Multimedia Learning in an Interactive Self-Explaining Environment: What Works in the Design of Agent-Based Microworlds? *Journal of Educational Psychology, 95*(4), 806-813. doi:10.1037/0022-0663.95.4.806

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model - Noncentrality and goodness of fit. *Psychological Bulletin, 107*, 247-255. doi: 10.1037/0033-2909.107.2.247

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.

Muthén, L. K., & Muthén, B. O. (1998-2013). *Mplus user's guide.* (Sixth ed.). Los Angeles, CA: Muthén & Muthén.

Naisbitt, J., & Cracknell, J. (1984). *Megatrends: Ten new directions transforming our lives*.

Nickerson, R. S. (1988). Technology in education in 2020: Thinking about the not-distant future. *Technology in education: Looking toward, 2020*, 1-24.

Pedhazur, E. J., & Pedhazur, S. L. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, N. J.: Erlbaum.

Poppenk, J., Köhler, S., & Moscovitch, M. (2010). Revisiting the Novelty Effect: When Familiarity, Not Novelty, Enhances Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 5.

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

Salaberry, M. R. (2001). The use of technology for second language learning and teaching: A retrospective. *The Modern Language Journal*, *85*, 39-56.

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

Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist*. doi:10.1080/00461520.1991.9653136

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.

Stickler, U., & Hampel, R. (2015). *Transforming Teaching: New Skills for Online Language Learning Spaces*. London: Palgrave Macmillan UK.

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

Veletsianos, G. (2010). Contextually relevant pedagogical agents: Visual appearance, stereotypes, and first impressions and their impact on learning. *Computers & Education, 55*, 576-585. doi:http://dx.doi.org/10.1016/j.compedu.2010.02.019

Veletsianos, G., & Miller, C. (2008). Conversing with pedagogical agents: A phenomenological exploration of interacting with digital entities. *British Journal of Educational Technology, 39*, 969-986.

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

Wikipedia. (n.d.). https://en.wikipedia.org/wiki/Cleverbot. Retrived on January 22, 2017.

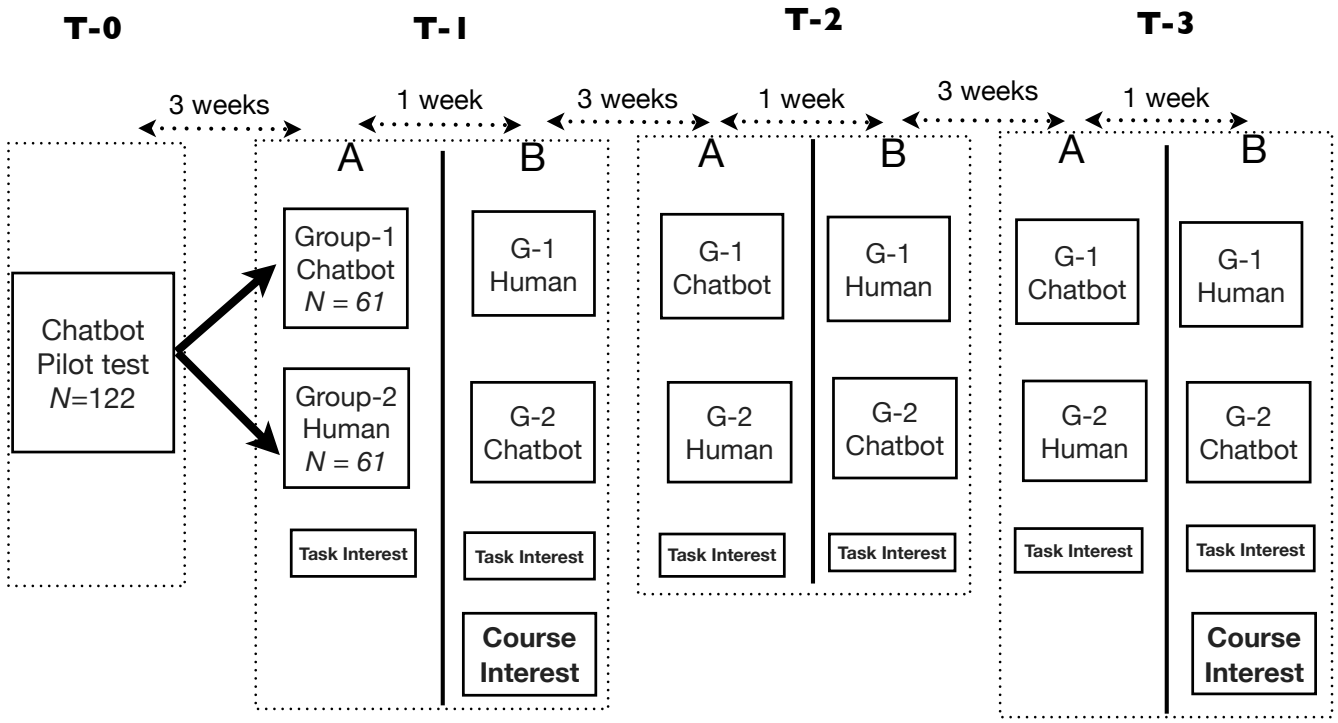

Figure 1**.** Research design showing sequence of task interest and course interest measures with counterbalanced administration of Chatbot and Human partner conditions at T1, T2 and T3.
Note: n=61 for each of the two groups

Table 1.
Latent Correlations and Descriptive Statistics for Task Interest and Course Interest Measures

		1	2	3	4	5	6	7	8
Task Interest									
1	Human T1								
2	Chatbot T1	.77							
3	Human T2	.77	.74						
4	Chatbot T2	.75	.74	.85					
5	Human T3	.75	.79	.85	.84				
6	Chatbot T3	.74	.74	.70	.74	.75			
Course Interest									
7	Course T1	.69	.56	.59	.59	.58	.55		
8	Course T3	.52	.39	.53	.46	.46	.55	.70	
	Mean	3.85	3.75	3.72	3.35	3.75	3.37	4.10	3.96
	SD	1.04	1.17	1.19	1.35	1.22	1.28	1.06	1.06
	Cronbach's Alpha	.93	.97	.96	.96	.96	.96	.93	.94

Table 2.
Mean Task Interest Scores for Three Time and Two Partner Conditions

Time	Chatbot Partner	Human Partner	Total
T-1	3.75[a]	3.85[a]	3.80 [a]
T-2	3.35[b]	3.72 [a]	3.53 [b]
T-3	3.37[b]	3.75 [a]	3.56 [b]
Total	3.54[b]	3.81 [a]	

Note: All pairwise test account for multiple comparisons: Bonferroni.
Means with the same superscript are not significantly different.

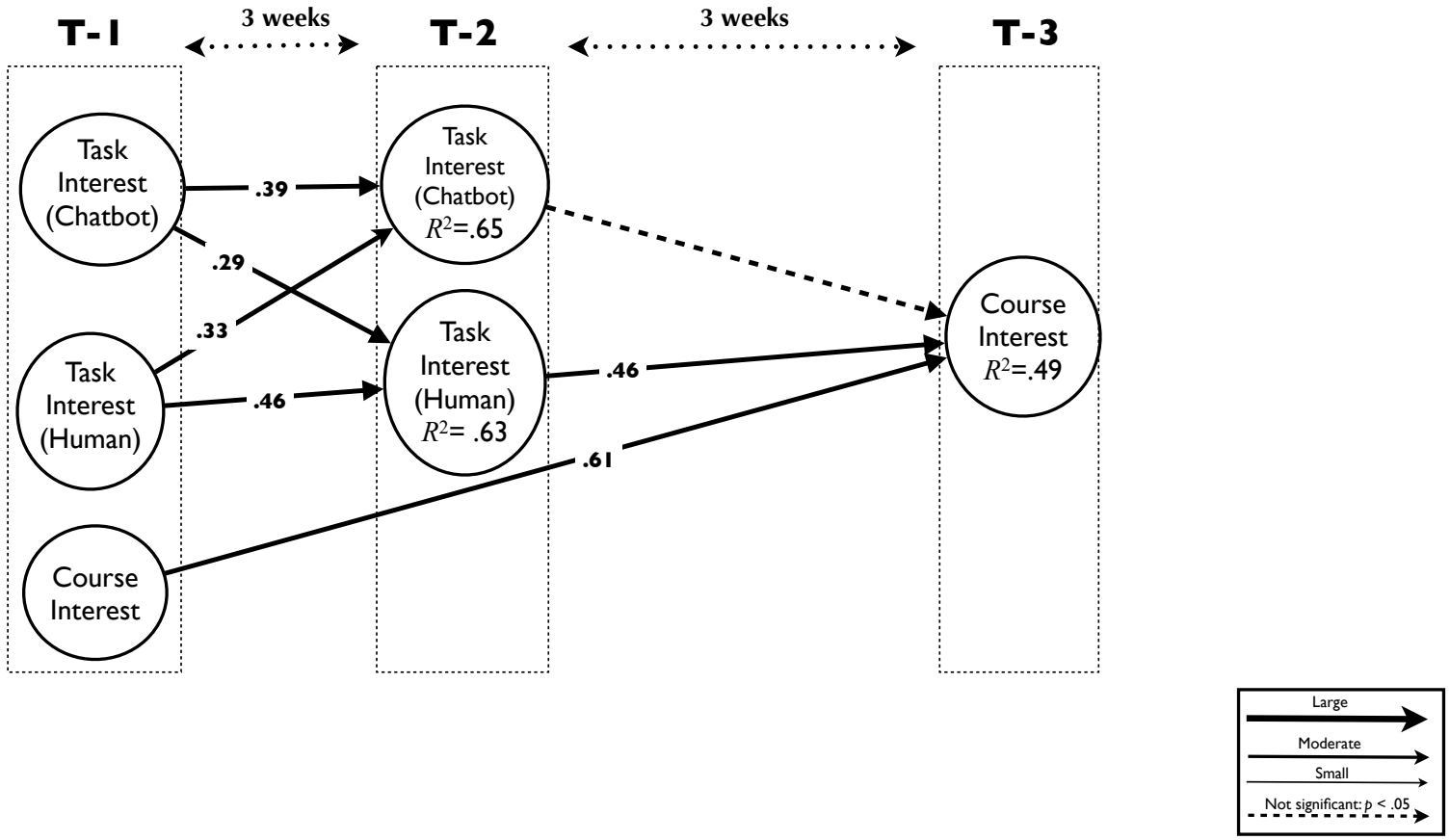

Figure 2. Significant predictive coefficients between task interest under Chatbot and Human partner conditions as they predict to T3 course interest.
Note: All coefficients are *ßs;* The dashed arrow represents the tested path that was not significant ($p < .05$).