Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Using Qualitative Data from Targeted Interviews to Inform Rapid AIED Development

Jaclyn OCUMPAUGH[a*], Stephen HUTT[a], Juliana Ma. Alexandra L. ANDRES[a], Ryan S. BAKER[a], Gautam BISWAS[b], Nigel BOSCH[c], Luc PAQUETTE[c] and Anabil MUNSHI[b]
[a] *University of Pennsylvania, USA*
[b] *Vanderbilt University, USA*
[c] *University of Illinois at Urbana Champaign, USA*
*ojaclyn@upenn.edu

Abstract: This paper examines how interviews with students—at critical moments of the learning process—may be leveraged to improve the design of educational software. Specifically, we discuss iterative work to improve the design of a pedagogical agent in the Betty's Brain learning environment, Mr. Davis. Students interacted with the pedagogical agent in Betty's Brain during two separate studies, two months apart. During study one, qualitative interviews were prompted by student actions within the system and theoretically aligned sequences of educationally relevant affective states (as detected by previously validated models). Facilitaed by an app called the Quick Red Fox (QRF), these *in situ* interviews were then used to identify ways to rapidly improve Mr. Davis' design, investigated in study two. Results indicate that changes designed to make Mr. Davis more empathetic correlate with improved learning outcomes. We also discuss the potential for rapidly collected qualitative data in future developments.

Keywords: Affective Computing, Learning by Teaching, Data-Driven Design, Pedagogical Agents

1. Introduction

Computer-based learning systems that seek to imitate human tutoring are faced with considerable challenges related to the design of their virtual pedagogical agents. Human tutors can dynamically adjust their interactions with students, accommodating differences in age, personality, cultural expectations, as well as moment-to-moment differences in affect or understanding. In contrast, virtual pedagogical agents are far less flexible, typically relying on a predefined set of tutorial actions or phrases (Veletsianos and Russell, 2014). Since it is not always possible to anticipate how students will interact with an agent (Kenkre and Murthy 2016), researchers need a way to quickly identify potential changes in student behavior and suggest improvements.

Pedagogical agents offer great potential for improving learning and supporting learners. Early work has shown that virtual agents can support increases in performance (Dumdumaya et al. 2017), provide beneficial social interactions (Doering, Veletsianos, and Yerasimou 2008; Kim and Wei 2011), and foster motivation and engagement (Kim and Wei 2011; Lusk and Atkinson 2007). However, some aspects of what makes a good interaction with a virtual agent is still an open question. Research has sought to calibrate pedagogical agents based on politeness, with many students responding well to more polite approaches (Graesser 2011; Person et al. 1995; Tynan 2005), but there have sometimes been conflicting results. Some studies have found that learning decreases when students perceive their tutors as irritating or offensive (De Angeli and Brahnam 2008). However, this pattern is not universal; other work has also shown that in some contexts, students respond well to so-called "rude tutors" who are designed to offer sarcastic responses even to struggling students (e.g., Graesser, 2011; Ogan, Finkelstein, Mayfield, et al. 2012).

Conflicting responses to politeness strategies are perhaps not surprising. Given the wide range of the functions of politeness (and impoliteness) that human tutors use (see review in Ogan, Finkelstein, Walker, et al. 2012), cultural, contextual, and developmental differences in student populations would likely moderate these findings (Savard and Mizoguchi 2019). Cultural differences are, unfortunately,

more difficult to tease out from the current literature than developmental differences (Paquette et al. 2020), but given findings that show high and low-knowledge learners respond differently to politeness strategies (D'Mello and Graesser 2013), developmental differences seem a likely explanation.

Thus far, most of the research on politeness and empathy has taken place with older adolescents and adult learners (Ogan, Finkelstein, Walker, et al. 2012; Rodrigo et al. 2013; Wang et al. 2012) (D'Mello and Graesser 2013). Less is known about the politeness practices of primary learners (ages 5-11), where students might be more likely to expect more empathetic pedagogical strategies.

This paper explores younger (6th grade) students' perception of two different versions of Mr. Davis, the mentor agent in the Betty's Brain learning system (Biswas, Segedy, and Bunchongchit 2016). Specifically, during *in situ* interviews with students, which were facilitated by an app called Quick Red Fox (QRF), we discovered that students were interpreting certain phrases being used by Mr. Davis as deliberately offensive (which was not the original intent of the designer), and unlike previous research on rude tutorial moves (e.g., Graesser 2011; Ogan, Finkelstein, Walker, et al. 2012), students were not responding well to these perceived slights. Accordingly, we followed up with further questions about what would make Mr. Davis feel more helpful. We then modified Mr. Davis's responses to make Mr. Davis seem more supportive. Critically, the insights occurred during initial reflection on the interviews, prior to more formal transcription and coding. This allowed for changes to be made before students used the software again two months later (for a new learning topic). Following this second session, students were surveyed about their current perceptions of Mr. Davis as well as their perceptions of how difficult the task was. In this paper, we report on what those changes are and then compare the perception of the two versions of Mr. Davis with learning gains as experienced by 99 middle schoolers across two studies.

2. Betty's Brain

Betty's Brain uses a learning-by-teaching model (Biswas et al. 2004), where students must teach a virtual agent named Betty by creating a causal map of a scientific process (e.g., climate change). Betty demonstrates her "learning" by taking quizzes that are graded by a pedagogical agent, Mr. Davis. As students construct Betty's map, they must navigate various learning resources, including hypermedia resources and a teaching manual that explains how to represent causal reasoning. The system is open-ended, with students choosing when to consult hypermedia resources, when to devote time expanding their causal map, and when to test it via the quizzes (Biswas et al. 2016).

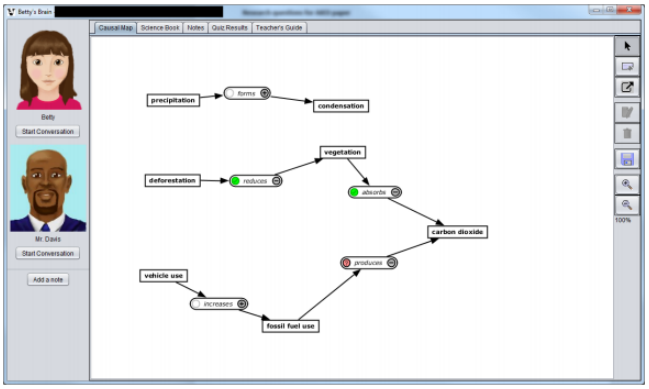

Figure 1. A partial casual map in Betty's Brain

Each time the students test their causal map (by testing Betty), Mr. Davis interacts with the students, giving them high-level suggestions on how they may improve. However, students can interact with him at other times in the learning process, asking him questions or seeking help/tips regarding causal maps. Mr. Davis is described to the students as an experienced teacher and is designed to present a helpful and mentoring role as the students teach Betty.

3. Methods

3.1 Data

Data were collected over two studies with 99 6th-graders who used Betty's Brain as part of their normal science instruction in an urban public school in the United States of America. In Study 1 (December 2018), students used Betty's Brain to learn about climate change. During the study, short interviews (usually 1-2 minutes) were repeatedly conducted, with researchers asking students to briefly pause their activity to talk about their work (more details in section 3.2). In Study 2 (February 2019), the same group of students interacted with Betty's Brain again and participated in a similar interview process as they learned and created a causal model for a new topic: thermoregulation.

Learning gains were assessed during both studies using pre- and post-test measures. For each study, the pre-test about the content (i.e., climate change or thermoregulation) knowledge was administered before students began working with the system. An identical post-test was administered at the end of the study. In this work, we characterize learning gains as post-test score – pre-test score.

3.2 Interviews

In both studies, interviews were dynamically prompted by real-time monitoring of affective and behavioral sequences (detected by previously validated models - see Jiang et al. 2018) that were being studied as part of a larger project on student affect and self-regulated learning (Bosch et al. 2021). In particular, key shifts in affect, such as *confusion → delight, confusion → frustration, frustration → engaged*, as well as *sustained* delight, were used to trigger interviews. Other interviews were prompted by specific behavioral sequences, including sequences that identified how effective the students' behaviors were. That is, actions like *editing the causal map* (as opposed to *opening a reading passage*) were further categorized by whether those edits were correct or incorrect. Thus, interviews were taking place during times that had been strategically identified to represent shifts in emotional states and key moments in the learning process.

Interviewers were directed to students experiencing one of these prompts by an Android-based app called Quick Red Fox (QRF), which collected metadata (deidentified student IDs and timestamps) while recording both the behavioral or affective sequence that triggered the interview and the interview itself. The interviewer then engaged the students with open-ended questions like "how are things going so far?" and "what strategies are you using to solve this problem?" Students were sometimes also asked about their intrinsic interest in science (e.g., "What is your favorite class?"), but in general, the interviewer let students guide the topics of interviews.

Though interviews were eventually manually transcribed and qualitatively analyzed (Bosch et al. 2021), this was not completed until after both rounds of data collection. In the interim, interviewers summarized their perceptions about trends and issues to the broader research and development team.

In the second study (after changes described in the next section were implemented), interviews were triggered and conducted using the same methods, but interviewers added explicit questions about the changes to the design of Mr. Davis.

3.3 Changes to Mr. Davis

During the Study 1 interviews, two scripts employed by Mr. Davis repeatedly emerged as particularly upsetting to the students, both of which involved his text beginning with the word "Hmph." Students consistently reported that they found Mr. Davis's interactions abrupt, frustrating, and, in some cases, distracting. Interviewers, therefore, asked students for suggestions on how to rewrite the scripts, including suggesting simply changing "Hmph" to "Hmm," which students found contemplative and less confrontational. Students also suggested that having Mr. Davis provide hints or encouragement would further improve his design. These design suggestions were reported to the development team, who made the appropriate changes to the conversations by incorporating ten new interactions and modifying two existing interactions. Three examples are shown in in Table 1.

Table 1. Breakdown of modifications to Mr. Davis

Type of Change	Example	N
Add new feedback with hints/guidance	"Hey, from the quiz results, it looks like Betty may have some incorrect links on her map. You can mark those links as could be wrong in your map. Do you want to know more about marking links as could be wrong?"	7
Add new feedback with encouragement	"Looks like you are doing a good job teaching correct causal links to Betty! Make sure that you check her progress from time to time by asking her to take a quiz."	3
Representational change for politeness	"Hmph" → "Hmm"	2

3.4 Surveys of Helpfulness and Difficulty

At the end of Study 2, students were asked to rate the helpfulness of Mr. Davis during each scenario (reflecting on both Study 1 and Study 2) using a single-item Likert measure from 1 (strongly dislike) to 5 (strongly like). Students were also asked to rate the difficulty of each scenario (also using a single item Likert measure).

4. Results

We examine the effectiveness of the changes to Mr. Davis by comparing the Likert scales on the perceptions about him between the two scenarios. We then explore the relationship between these results in terms of learning and student perceptions of difficulty.

4.1 Perceptions of Mr. Davis

We first considered if the changes to Mr. Davis improved student perception of him. A paired samples Wilcoxon signed-ranks test showed a statistically significant improvement in how helpful Mr. Davis was perceived as being across the two scenarios (Study 1: $M = 2.10$, $SD = 1.04$, Study 2: $M = 2.83$, $SD = 1.37$; $Z=.43$, $p < 0.001$). This aligns with the qualitative reports given during student interviews in the second study. A histogram of student responses regarding Mr. Davis is shown in Figure 2.

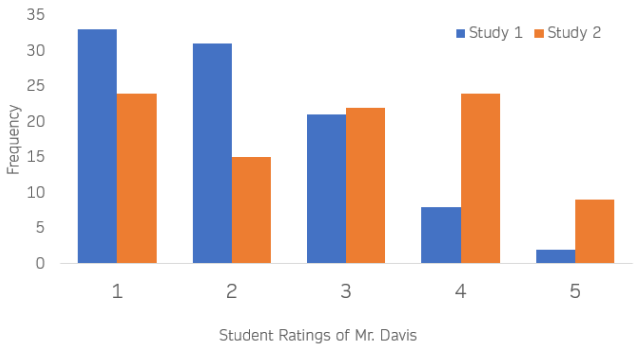

Figure 2. Bar Chart of Student Scores of Mr. Davis across both studies (1 = Strong Dislike, 5 = Strong Like)

4.2 Mr. Davis and Learning Gains

We next examined how students' perceptions of Mr. Davis were linked to their learning gains within Betty's Brain. For each student, we calculated the difference between their two survey responses (Likert score for Study 2 minus Likert score for Study 1) with a positive score indicating that students felt Mr.

Davis had improved. Through Spearman correlations, we found that improvement perception of Mr. Davis was positively correlated with learning gains in the second topic (rho = .37, $p < 0.01$).

To further examine if this result was an artifact of student self-efficacy (e.g. students being more likely to perceive Mr. Davis as helpful if they are doing well in the Betty's Brain), we examined the learning gains in the first scenario (prior to the changes to Mr. Davis) and observed no significant correlation to improvements in Mr. Davis (rho = .08, p = .46), reinforcing our conclusion that the changes to Mr. Davis had a positive impact on students' experiences. However, it should be noted that correlation is not causation, and other factors may have had an impact (see discussion).

4.3 Mr. Davis and Perceptions of Difficulty

Finally, we examined the difference in students' perceived difficulty of the two topics. We found that the change in students' perception of Mr. Davis was negatively correlated with change in perceived difficulty (*rho* = -.31, $p < 0.01$), indicating that students who found Mr. Davis more helpful the second time than the first time also found the second scenario easier than the previous scenario. There is insufficient evidence to be certain of causality, but this evidence is compatible with the idea that the students felt more supported during their second encounter with the system.

5. Discussion and Conclusions

Results show that the changes implemented between studies 1 and 2 led to more favorable interpretations of Mr. Davis. Specifically, we show a considerable increase in the perception of Mr. Davis as being helpful. Students' improved perceptions of Mr. Davis were also correlated with an improvement in learning gains and a reduction in perceived difficulty.

A limitation of this work is that we have not considered how additional external factors such as student familiarity with the system and topic familiarity may have impacted these perceptions due to scope. While our survey measures consider Mr. Davis in isolation, it is likely that more general perceptions about the system also have an impact here. More research is needed to determine how students' perceptions of Mr. Davis interact with other parts of the learning process (such as broader self-regulation), as is more research on what factors impact whether students prefer more polite and supportive pedagogical agents. However, this study demonstrates how even minor changes at the pragmatics level of conversation (i.e., changes that effect the perception of politeness and intent) may impact students' interactions with a learning system, suggesting that researchers should explore how politeness may interact with other parts of the design of computer-based learning environments.

This study also demonstrates the importance of exploring new methodologies for collecting feedback on student-tutor interactions. In particular, it suggests that rapidly collected qualitative data, such as those collected in the interviews prompted by the QRF app, might be useful for rapid design iteration and improvement in real-world settings. This is especially true in situations when implementing in a new context (e.g., with students in a new age group, or implementing in a new language). We often rely on quantitative data or formally coded interviews to drive design work, but this is often costly and time-consuming, meaning that relatively minor changes to educational software can take considerable time. This work presents a useful substitute for laboratory-based usability studies when those are impractical to conduct or may change students' attitudes towards the learning experience. While it sometimes difficult to anticipate what tutorial moves students will respond well to and exactly how to write tutorial dialogue, it can be very easy for a human interviewer to identify when a student responds poorly to a tutor's responses, allowing for timely adjustments to correct course.

Acknowledgments

This work was supported by NSF #DRL-1561567.

References

De Angeli, Antonella, and Sheryl Brahnam. 2008. "I Hate You! Disinhibition with Virtual Partners." *Interacting with Computers* 20(3):302–10.

Biswas, Gautam, Krittaya Leelawong, Kadira Belynne, Daniel Schwartz, and Joan Davis. 2004. "Incorporating Self Regulated Learning Techniques into Learning by Teaching Environments." in *The Twenty Sixth Annual Meeting of the Cognitive Science Society*.

Biswas, Gautam, James R. Segedy, and Kritya Bunchongchit. 2016. "From Design to Implementation to Practice a Learning by Teaching System: Betty's Brain." *International Journal of Artificial Intelligence in Education*.

Bosch, Nigel, Y. Zhang, Luc Paquette, Ryan Shaun Baker, Jaclyn Ocumpaugh, and Gautam Biswas. 2021. "Students' Verbalized Metacognition during Computerized Learning." P. 12 in *ACM SIGCHI: Computer-Human Interaction*. Association for Computing Machinery.

Cohen, J. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Taylor & Francis.

D'Mello, Sidney K., and Art Graesser. 2013. "AutoTutor and Affective Autotutor: Learning by Talking with Cognitively and Emotionally Intelligent Computers That Talk Back." *ACM Transactions on Interactive Intelligent Systems* 2(4).

Doering, Aaron, George Veletsianos, and Theano Yerasimou. 2008. "Conversational Agents and Their Longitudinal Affordances on Communication and Interaction." *Journal of Interactive Learning Research* 19(2):251–70.

Dumdumaya, Cristina E., Michelle P. Banawan, Ma Rodrigo, T. Mercedes, Amy Ogan, Evelyn Yarzebinski, and Noboru Matsuda. 2017. "Investigating the Effects of Cognitive and Metacognitive Scaffolding on Learners Using a Learning by Teaching Environment." in *International Conference on Computers in Education (ICCE)*.

Graesser, Arthur C. 2011. "Learning, Thinking, and Emoting With Discourse Technologies." *American Psychologist* 66(8):746.

Jiang, Yang, Nigel Bosch, Ryan S. Baker, Luc Paquette, Jaclyn Ocumpaugh, Juliana Ma Alexandra L. Andres, Allison L. Moore, and Gautam Biswas. 2018. "Expert Feature-Engineering vs. Deep Neural Networks: Which Is Better for Sensor-Free Affect Detection?" Pp. 198–211 in *Artificial Intelligence in Education*.

Kenkre, A., and S. Murthy. 2016. "Students Learning Paths in Developing Micro-Macro Thinking: Productive Actions for Exploration in MIC-O-MAP Learning Environment." in *International Conference on Computers in Education (ICCE)*.

Kim, Yanghee, and Quan Wei. 2011. "The Impact of Learner Attributes and Learner Choice in an Agent-Based Environment." *Computers & Education* 56(2):505–14.

Lusk, Mary Margaret, and Robert K. Atkinson. 2007. "Animated Pedagogical Agents: Does Their Degree of Embodiment Impact Learning from Static or Animated Worked Examples?" *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 21(6):747–64.

Ogan, Amy, Samantha Finkelstein, Elijah Mayfield, Claudia D'Adamo, Noboru Matsuda, and Justine Cassell. 2012. "'Oh, Dear Stacy!' Social Interaction, Elaboration, and Learning with Teachable Agents." Pp. 39–48 in *Conference on Human Factors in Computing Systems*.

Ogan, Amy, Samantha Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. 2012. "Rudeness and Rapport: Insults and Learning Gains in Peer Tutoring." Pp. 11–21 in *International Conference on Intelligent Tutoring Systems*.

Paquette, Luc, Jaclyn Ocumpaugh, Ziyue Li, Alexandra Andres, Ryan Baker, and others. 2020. "Who's Learning? Using Demographics in EDM Research." *JEDM| Journal of Educational Data Mining* 12(3):1–30.

Person, Natalie K., Roger J. Kreuz, Rolf A. Zwaan, and Arthur C. Graesser. 1995. "Pragmatics and Pedagogy: Conversational Rules and Politeness Strategies May Inhibit Effective Tutoring." *Cognition and Instruction* 13(2):161–69.

Rodrigo, Ma Mercedes T., R. Geli, Aaron Ong, G. Vitug, Rex Bringula, R. Basa, and N. Matsuda. 2013. "Exploring the Implications of Tutor Negativity towards a Synthetic Agent in a Learning-by-Teaching Environment." *Philippine Comput. J* 8:15–20.

Savard, Isabelle, and Riichiro Mizoguchi. 2019. "Context or Culture: What Is the Difference?" *Research and Practice in Technology Enhanced Learning* 14(1):23.

Tynan, Renee. 2005. "The Effects of Threat Sensitivity and Face Giving on Dyadic Psychological Safety and Upward Communication." *Journal of Applied Social Psychology* 35(2):223–47.

Veletsianos, George, and Gregory S. Russell. 2014. "Pedagogical Agents." Pp. 759–69 in *Handbook of research on educational communications and technology*. Springer.

Wang, William Yang, Samantha Finkelstein, Amy Ogan, Alan W. Black, and Justine Cassell. 2012. "'love Ya, Jerkface': Using Sparse Log-Linear Models to Build Positive (and Impolite) Relationships with Teens." Pp. 20–29 in *13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*.