



## Feasibility and usability of MentorPal, a framework for rapid development of virtual mentors

Benjamin D. Nye, Dan M. Davis, Sanad Z. Rizvi, Kayla Carr, William Swartout, Raj Thacker & Kenneth Shaw

**To cite this article:** Benjamin D. Nye, Dan M. Davis, Sanad Z. Rizvi, Kayla Carr, William Swartout, Raj Thacker & Kenneth Shaw (2021) Feasibility and usability of MentorPal, a framework for rapid development of virtual mentors, *Journal of Research on Technology in Education*, 53:1, 21-43, DOI: [10.1080/15391523.2020.1771640](https://doi.org/10.1080/15391523.2020.1771640)

**To link to this article:** <https://doi.org/10.1080/15391523.2020.1771640>



Published online: 22 Jul 2020.



Submit your article to this journal [↗](#)



Article views: 462



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 7 View citing articles [↗](#)



## Feasibility and usability of MentorPal, a framework for rapid development of virtual mentors

Benjamin D. Nye<sup>a</sup> , Dan M. Davis<sup>a</sup>, Sanad Z. Rizvi<sup>a</sup>, Kayla Carr<sup>a</sup>, William Swartout<sup>a</sup>, Raj Thacker<sup>a</sup>, and Kenneth Shaw<sup>b</sup>

<sup>a</sup>Institute for Creative Technologies, University of Southern California, Los Angeles, California, USA; <sup>b</sup>Georgia Institute of Technology, Atlanta, Georgia, USA

### ABSTRACT

One-on-one mentoring is effective for helping novices with career development. However, traditional mentoring scales poorly. To address this problem, MentorPal emulates conversations with a panel of virtual mentors based on recordings of real STEM professionals. Students freely ask questions as they might in a career fair, while machine learning algorithms respond with best-match answers. MentorPal is researching rapid development of new virtual mentors, where training data will be sparse. In a usability study, 31 high school students reported (a) increased career knowledge and confidence, (b) positive ease-of-use, and that (c) mentors were helpful (87%) but seldom covered their preferred career (29%). These results demonstrate feasibility for virtual mentoring, but efficacy studies are needed to evaluate its impact, particularly for groups with limited STEM opportunities.

### ARTICLE HISTORY

Received: 1 November 2019  
Revised 10 April 2020  
Accepted 16 May 2020

### KEYWORDS

Virtual agents; mentoring; natural language processing; artificial intelligence; STEM outreach

Projections indicate that up to 2,000,000 Science, Technology, Engineering, and Mathematics (STEM) positions may be unfilled by 2025, due to a lack of qualified workers for these new jobs (Wellener et al., 2018). STEM careers grew much faster (24%), than the non-STEM areas over the previous decade (Noonan, 2017) and STEM occupations have been projected to grow by 9% vs. 6% overall (40% faster) from 2014 to 2024. Unfortunately, students have minimal understanding about what professionals in STEM fields do on-the-job, what education is needed, and even what STEM fields exist (Bieber et al., 2005). Together, these barriers discourage many students from pursuing STEM by high school (Wang, 2013).

Career fairs, informational interviews, alumni networks, and student mentoring centers play important outreach roles, particularly for historically underrepresented populations that often lack access to role models and knowledge about STEM careers. These interactions increase motivation, engagement, and career self-efficacy (Herman, 2010; Krieschok, 1998; Stoecker et al., 2013). However, three structural challenges limit mentor outreach. First, scalability is limited by mentors' availability (time, location). Second, each center has only a limited set of local mentors and who may not be in the fastest-growing careers. Further, such mentorship programs may reproduce entrenched STEM demographics (Lehming et al., 2013). Finally, transfers, job changes, retirement, and other life events can cause mentor attrition.

Virtual agents offer the potential to amplify the knowledge and insights from compelling mentors and role models. This article describes the design and a usability study for MentorPal, a

virtual agent framework designed to scale up mentoring and career insights on a broad scale (Nye et al., 2017). MentorPal rapidly develops virtual mentors based on recordings of unscripted, live mentors answering questions about their career path. Students freely ask questions, each mentors' machine learning predictive algorithm attempts to select and play back the best response recording from that mentor, as if they were holding a live video chat. Insights from formative testing led to multiple design iterations (Breck et al., 2018; Nye et al., 2017), including the strategy of using a MentorPanel with four mentors at a time. By modeling individual STEM professionals, this approach shares the career trajectories of individuals already in that job (Bandura, 1986). If this approach is feasible, it could amplify under-represented mentors in STEM (e.g., minorities and women) by capturing a diverse set of mentors and careers.

To achieve this vision, virtual mentors must be fast and affordable to develop. As is evident in reviews of pedagogical agent design (Johnson & Lester, 2016; Schroeder et al., 2013), substantial effort is required to build an intelligent virtual agent—typically only one or two per multi-year project. On the converse, while they are only momentarily available, career fairs still have dozens of career professionals. For a virtual career fair, it would be ideal to represent many careers, with professionals of different genders, ethnicities, ages, socio-economic backgrounds, and geographic origins. Populating such a fair implies hundreds of virtual mentors.

To approach this challenge, a new technique for rapidly developing virtual mentors is essential. However, substantial constraints are required for rapid mentor development: building mentors must be automatable (no essential human steps to turn videos and transcripts into a mentor), mentors must record only a modest set of questions (25% the size of known high-quality agents; Artstein et al., 2014), and recordings must have only basic video production values (webcam with unscripted, unedited answers). Due to these constraints, it was uncertain that these mentors would be accepted by users (i.e., if they were feasible in practice). The primary research questions for the usability study were to confirm user acceptance of the mentors:

(Q1) Expected Benefit: Will students report positive expected benefit (knowledge, confidence) from virtual mentors?

(Q2) Ease of Use: Will students report positive ease-of-use when interacting with virtual mentors?

(Q3) Mentor Coverage: Will a panel of four virtual mentors be sufficient to identify at least one helpful mentor to interview and gain knowledge about careers?

Data was also collected to study secondary research areas, specifically the impact of this brief self-guided virtual mentor intervention for (a) changes in their interest in specific careers represented in the panel and (b) changes in their attitudes about types of STEM careers more generally. Due to the sample size and limited number of virtual mentors, these findings are formative: while they indicate shifts in career attitudes, the impact and value of these shifts is not conclusive.

## Theoretical background

### *Career mentors: roles and characteristics*

Career mentors are not interchangeable: they have different characteristics that make them better suited to specific mentees. In a survey of mentors, Kraiger et al. (2019) found that in short term mentoring (0–3 months), mentors focused strongly on the mentoring relationship, the mentee's attitudes (e.g., accountability, confidence), and building mentee competence (e.g., skills, knowledge). In longer term settings, mentors increased focus on the broader career context (e.g., professional networks, career progression). While not all of these factors are relevant to virtual interactions, MentorPal targets short-term mentoring goals for building: (a) a positive

connection/relationship, (b) career attitudes (confidence), and (c) useful strategies and knowledge (competence).

Mentee's perceptions of the effectiveness of their mentors have also been found to be influenced by their interpersonal characteristics. In a meta-analysis of 173 studies ( $N=40,737$ ), Eby et al. (2013) found that perceived instrumental support was most-influenced by a mentor's deep similarity ( $\rho = .38$ ; e.g., attitudes) and experiential similarity ( $\rho = .21$ ; e.g., life experiences). By comparison, shallower factors were not influential, such as surface-level similarity ( $\rho = .03$ ; race and gender) and mentor human capital ( $\rho = -.11$ ; e.g., education level). These findings were similar for mentees' perceived psycho-social support, except that deep similarity showed even stronger influence while experiential similarity became trivial ( $\rho = .03$ ). Thinking about a career as a trajectory: deep similarity can imply similar goals and behavioral standards (strategies), while experiential similarity should increase the likelihood that a mentor knows a strategy relevant to the mentee (e.g., socio-economic constraints).

However, surface-level similarity still has notable direct and indirect influences. First, demographic similarity is visible at-a-glance, so it may increase the likelihood of initiating a mentoring interaction. Second, race and gender are both influences on experiential similarity, such as differential responses to negotiating strategies based on gender expectations (Amanatullah & Morris, 2010). Finally, lack of representation for one's demographic in a STEM career intervention may result in the perception of "not for people like me" and discourage engagement (Macdonald, 2014).

### ***Virtual humans as mentors***

Synthetic, animated, virtual agents have been particularly effective at communicating information to a wide range of people using interactive dialog (Baylor, 2009; Baylor & Kim, 2016; Bickmore et al., 2010; Swartout et al., 2010; 2013; Traum et al., 2012). Often, people self-disclose more information to agents, since they do not feel judged (Gratch et al., 2014). This phenomenon may be valuable for interaction with virtual mentors because it indicates that people may ask a virtual mentor questions they would not ask live mentors.

Substantial research has investigated how agent appearance, behavior, interactions and voice prosody impact the role they are intended to fulfill (Kim & Baylor, 2016; Schroeder et al., 2013). Baylor and Kim (2016) showed that a mentor agent with both motivational and expertise-based interactions produced higher self-efficacy and learning. Moreover, it found that the agent's appearance impacted perceptions of the level of expertise and motivation to interact. Further research has shown that groups who feel under-represented in a domain benefit most from virtual mentors with similar demographics (e.g., pre-teen girls in mathematics; Arroyo et al., 2011).

Finally, our work was inspired by research to preserve the ability to interact with Holocaust survivors, called New Dimensions in Testimony (NDT; Traum et al., 2015). In NDT, maintaining authenticity was critical, so agents were embodied through video-recordings of real Holocaust survivors, answering the questions that visitors ask at Holocaust museums. Authenticity is essential for mentoring (Fries-Britt & Snider, 2015), which led MentorPal to focus on video-recorded mentors. As such, MentorPal gives up direct control of the content that a synthetic agent offers, but the agents gain authenticity because users see a real person, who can speak meaningfully about their own experiences (i.e., experiential similarity).

### ***MentorPal design influences***

The specific interaction that this system models is a 5–30-min informational interview with a professional, such as at a career fair. During an informational career interview, a student asks questions about a mentor's job to get a better understanding of how well that career fits their interests and goals. MentorPal leveraged approaches from two earlier projects: New Dimensions



**Figure 1.** NDT at the Illinois Holocaust Museum.

in Testimony (NDT) and the Personal Assistant for Life-Long Learning (PAL3). The NDT project, which is intended to maintain and share the experiences of Holocaust survivors (Figure 1), serves as a model for building systems to communicate personal life experiences (or career experiences).

In NDT, recorded answers and questions were transcribed, and a predictive algorithm was developed based on the question-answer pairs, using machine learning. This algorithm predicts the words that are likely to appear in the answer (Traum et al., 2015). Artstein et al. (2014) found that about 1,400 recorded answers could answer 70% of questions in an open-ended conversation, and iterative improvement allowed fewer than 2,000 answers to cover 95% of questions in a highly-natural conversation.

The user interface design for MentorPal was developed inside the PAL3 framework (Figure 2), which controlled the overall user flow (e.g., account creation, logging). The PAL3 project (Hampton et al., 2018; Swartout et al., 2016) is an adaptive learning platform, with an interactive learning assistant (Pal). The long-term goal of PAL3 is to track a learner's progress and provide personalized learning recommendations from a library of learning resources based on performance and career goals. Within this framework, MentorPal was integrated as a learning resource that could be recommended to users (and in this study, users navigated to it directly).

As compared to NDT, we hypothesized that by focusing more tightly on career-relevant dialog, effective conversation should be possible despite a smaller answer set. One goal of this project is to identify strategies that maintain effective interactions, while lowering costs closer to a point where it might be possible to model hundreds of mentors, but still retaining the engaging feeling of interacting with a real-life person.

To that end, MentorPal was guided by a set of three design principles aligned to the goal of cost-effective, wide-scale use of virtual mentors:

*Conversational Flow:* Rapid, real-time responses to user questions, including interrupting mentors, and mentors considering prior questions when responding.

*Video-Chat Authenticity:* Target a level of visual and audio fidelity emulating a high-quality videoconference (e.g., Skype or Google Hangouts) with the mentor.

*Low-Cost:* Specialized equipment for recordings should be no more than \$250, re-usable, and use open source tools.



Figure 2. PAL3 (Tablet Version) Home Screen.

## MentorPal design

### *MentorPal user interface*

The MentorPal interface was modeled after video conference designs. Initially, MentorPal presented a single mentor (Figure 3) and only two elements: a video panel and a text box. Users submitted a question and the system classifier found the best-matching answer, which triggered a video of the real-life mentor answering (or a video of the mentor stating they lacked an answer). Based on internal feedback and later user testing (Breck et al., 2018), new features were added including pause, replay, a conversation transcript with replay functionality, subtitles, and voice input questions (speech to text).

However, after testing with students, it was identified that many high school students had trouble generating productive questions. In response, to help students get started with questions we conducted brief practice sessions, where experimenters modeled asking one or two questions. Dialog prompts by the mentor also helped slightly (e.g., “You can ask me about how I started my career, my time in Japan ...”). While this reduced its severity, some students (particularly younger teens) continued to struggle. As a result, an alternate interaction mode was designed.

To scaffold question asking, topic buttons were added that cycle through possible questions (shown beneath the video panel in Figure 3). Topic buttons do not ask the question, but they fill the text box with potential questions. Cycling through one question at a time helps a student explore the space of questions they *should* be asking and that they could submit. While not assessed in this study, topic buttons may have a distinct learning effect, since manually exploring the questions associated with each topic should help students frame which factors to consider in career choice. As a result, MentorPal has three options for input:

*Free Text:* Typing a question into the text box, such as a specific concern or a follow-up question about something a mentor said. Pedagogically, both free text and speech input support generativity (McDaniel et al., 2007).

*Speech Input:* Activating the microphone to input speech, which is converted to text input to submit.



*Topic Buttons:* Cycle through suggested questions in a topic of interest, shown in the order that researchers prioritized them. Pedagogically, these buttons scaffold question-asking (Graesser et al., 2014).

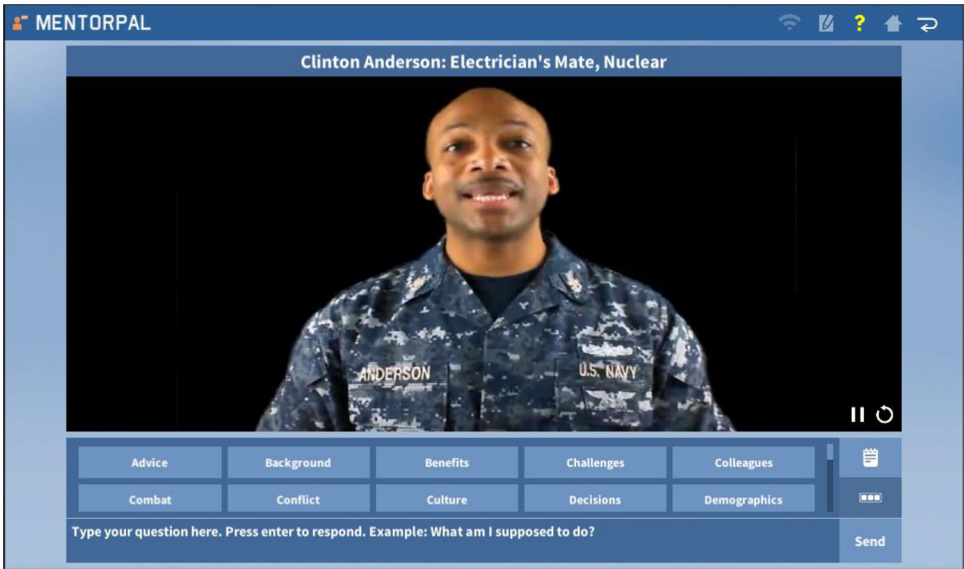


Figure 3. MentorPal Single-Mentor Interface.

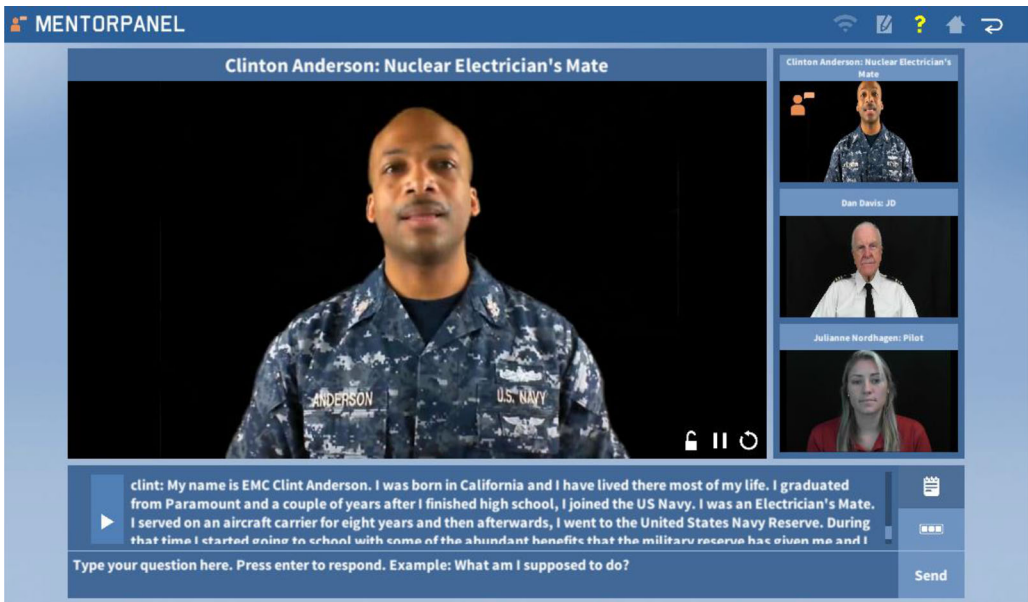


Figure 4. MentorPanel Interface.

Students reported the single mentor interface as useful (Breck et al., 2018), but this interface showed limitations after more mentors were recorded. The original concept assumed that a student would close a mentor if they wanted to talk to another. However, since MentorPal has no clear ending condition, students did not find it intuitive to switch mentors. MentorPanel was

designed to address this problem, which generalized the video conference metaphor to a group video chat, where users could ask questions and hear responses from different mentors (Figure 4). This feature was inspired by the US Naval Academy Alumni Mentoring Program ([www.usna.com/amp](http://www.usna.com/amp)) strategy for major career decisions, which seeks to find four mentors with different perspectives: two pairs with different choices (e.g., separating vs. staying in the Navy), where one of the pair was happy with their decision and one was less satisfied.

The advantages of MentorPanel quickly became clear. First, the panel approach offered an effective way to become familiar with at least four mentors and their points-of-view rather than just one. This seemed likely to give a better chance that users could find a career or individual who they preferred. Second, by adding a “lock” button, the MentorPanel could emulate a one-on-one chat. Third, by having multiple mentors offer their views, a student could compare and contrast different careers or viewpoints. Finally, the panel approach helps mitigate issues due to limited answer sets, since only some of the mentors need to answer a question for the system to feel conversational.

Feedback from early users also impacted the design process. Specifically, stakeholder sites expressed concern about using a downloadable application. This was due to the challenge of installing and debugging a local app and also due to the desire to reach students that might be difficult to reach for live mentoring, such as in rural areas. As a result, new iterations of MentorPal are now web-based (Figure 5).

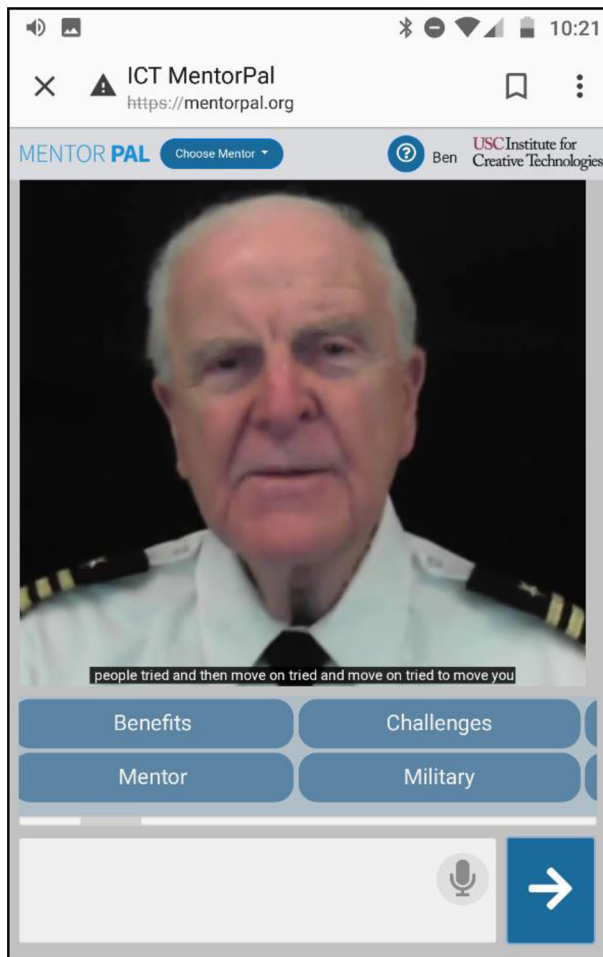


Figure 5. Web Based MentorPal (Mobile View).



**Table 1.** Mentor careers, demographics, and training set size (at time of recording).

Mentor	Career(s)	Career Stage	Demographics & Geographics	# Answers	# Q&A Training
Clint	- Nuclear Electrician's Mate Chief (EMC) - Computer Science	Mid	African American Male. In Japan & CA, and from CA, Flint, and Alabama.	381	2623
Dan	- High Performance Computing Research (Supercomputers) - Navy Cryptologist (CDR)	Late	White Male. In CA and from Colorado Springs.	365	2671
Carlos	- Marine Logistics (Civ) - Logistics (Warrant Off.)	Mid	Latin American Male. In South Carolina, from Philadelphia & New Haven.	307	1896
Julianne	- Systems Engineering - Naval Aviator (ENS)	Early	White Female. In CA and from Boca Raton, FL.	61	307

**Mentor recording content**

The MentorPal technology is only a medium: effective mentoring is determined by the quality of the engagement and advice from the human mentor. The full panel of mentors was selected to span a diverse set of backgrounds and career stages, as shown in [Table 1](#) (listed in order of recording). The primary inclusion criteria for mentors were that they had prior experience as mentors in their field. A secondary inclusion criterion was that mentors needed to be able to share their experiences using engaging stories.

It is important that students have opportunities to talk with mentors in different career fields and who they relate to in terms of life experiences. Recruiting for the panel targeted a balance of two women and two men. However, a female physicist selected was unable to continue, so a previously recorded male panelist was used for this panel. A second mentor also received transfer orders with only 1/5 of their questions recorded. However, it was identified that they could still work effectively as part of a panel. Despite the range of experiences and different careers in this prototype, the limitations of even four mentors were evident: only a small set of STEM fields were covered.

A goal of this project was to help students understand STEM careers relevant to both the Navy and to society overall, so mentors were selected with experience in Navy-relevant STEM careers. However, the goal of these recordings was not to advocate for a career but to give specific information about the realities of each career: both what makes their careers a good fit for someone and a bad fit for someone else. [Table 2](#) summarized the topics for each career. Some highlight career benefits (e.g., Growth) while others highlight difficulties or day-to-day headaches (e.g., Challenges). Nye et al. (2017) gives a more thorough overview of the mentor selection, recording, and content development process.

The primary pedagogical technique encouraged during recording was the use of narrative, such as brief anecdotes to make key information memorable and salient (Dahlstrom, 2014). Before each session, mentors were reminded to answer questions as if they were talking to high-school students. For responses with relatively low career information, mentors were advised to remain brief. While most answers were recorded only once, mentors were encouraged to re-record an answer to highlight an element that was particularly compelling, to improve conversation flow (e.g., long pauses, too long), or to allow mentors to re-phrase their answers.

As the goal of this project is also to reduce the time to produce a virtual mentor, [Table 3](#) breaks down estimates of our per-mentor content development time for the final mentor

**Table 2.** Examples of question topic categories, with example questions (see Nye et al., 2017 for the full set of 37 categories).

Category	Question
Background	Where did you grow up?
Benefits	How do the educational benefits of your work?
Challenges	What is your strategy for overcoming hardships?
Conflict	Tell me about a time when your priorities conflicted with that of management.
Culture	What do you find unique about your career field?
Ethics	Is sexual harassment common in the military?
Failure	What are the failures for which you are the most unhappy?
Mentor	Who are the best people to talk to when making a career choice?
Motivation	Why do you love your job?
STEM	What is something most people don't know about computer science?
Travel	How much travel did you do during active duty?

**Table 3.** Estimated hours to develop an additional virtual mentor.

Development step	Estimated hours (final mentor)
Video recording of the Mentor (360–400 questions)	18
Video postproduction (e.g., sound, lighting levels, manually marking timestamps to split video into clips)	16
Clean up automated transcripts	36
Revise mentor content based on real student questions	25
Administrative effort (scheduling, setup for 6 sessions)	15
<b>Total (Self-Recorded Mentor)</b>	<b>110</b>
<b>Total (One Interviewer + 20h)</b>	<b>130</b>
<b>Total (Two Interviewers + 40h)</b>	<b>150</b>

(Carlos). Assuming two researchers conduct the mentor interviews, the time to develop a mentor with a similar base question bank was approximately 150 h/mentor. These estimates are not for the first version of the mentor, but for a cleaned-up version where most transcription errors are corrected and some post-production is used to fix videos with sound or lighting problems. As such, the first prototype of an interactive mentor can be created in closer to 75 h of effort by two researchers and one mentor. This means mentors can be recorded and deployed within 2 weeks to produce approximately 15 h of mentoring responses, which is efficient for a virtual agent (e.g., about 10 h development per 1 h interactive content). Since this time, improvements have been integrated to increase automation (e.g., better automated transcriptions, audio smoothing).

### ***MentorPal dialog engine***

For the MentorPal system to emulate talking with a mentor, a dialog engine was required. This engine was actively revised to optimize multiple goals: response latency, classification quality, dialog moves (e.g., handling repeat questions), and system resources (e.g., available RAM; Kaimakis et al., 2018). The set of features used in the usability study included the ones that were most stable and useful, while certain experimental features were disabled.

An ensemble classifier was used for MentorPal, whose high-level data flow is outlined in Figure 6. An ensemble model was used because the classifier will be prone to the cold start problem for each new mentor. Specifically, while each mentor shared a large number of questions (75–80%), they also had mentor-specific questions based on their responses and backgrounds. Every mentor had a separate classifier trained with a data set that included standard question paraphrases, as well as paraphrases that were specific to that mentor.

The input stage of the model is relatively straightforward, with all inputs ultimately classified based on analysis of a text string, in the context or prior inputs. The Topic Button that generates questions suggested possible answers until all options were exhausted. Questions were not chosen

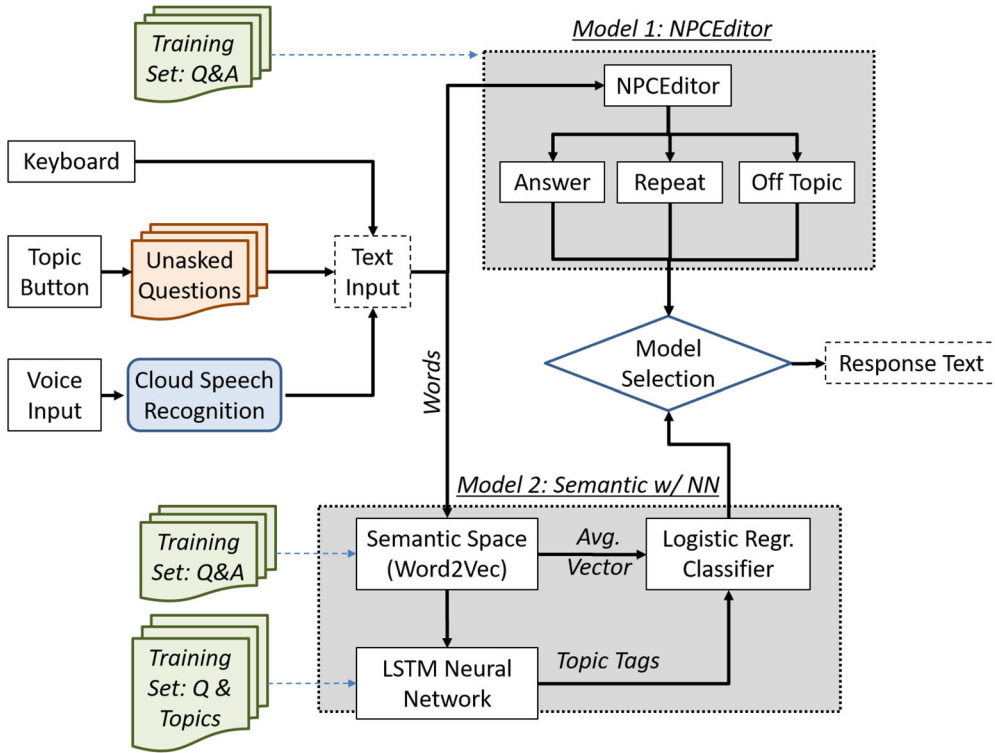


Figure 6. MentorPal Ensemble Classifier.

randomly; but were instead generated with an ordered list, so that more important questions could be prioritized.

The ensemble classifier consisted of two parts. The first was the established NPCEditor model (Leuski & Traum, 2011). NPCEditor was configured to report the best answer and meta-data indicating if the response was a repeat or if it was off-topic. NPCEditor was trained on a data table that listed a question, an answer, and a list of paraphrases that should produce the same answer. If the NPCEditor's confidence exceeded a threshold, the ensemble model accepted that response. As such, the ensemble acted as a fall-through backup so that questions not well-understood by the NPCEditor could have a second chance using Word Embedding and Neural Network model (Semantic w/NN).

This Semantic w/NN model used a Logistic Regression Classifier that relied on a combination of word embedding data (Word2Vec; Mikolov et al., 2013) and a long short-term memory neural network (LSTM; Abadi et al., 2016; Hochreiter & Schmidhuber, 1997). The Neural Network attempted to classify the topic(s) a question belonged to, based used received input ( $w$ ) as the average Word2Vec across the tokens for the question (words/phrases) and its output ( $t$ ) was a vector of binary indicators for which topics were relevant. The Logistic Regression classifier would then classify the best answer, with an input vector of average Word2Vec ( $w$ ) embeddings for the input question plus the neural-network assigned topics ( $t$ ).

### Dialog classifier optimization

The quality of the dialog system was tested formatively and revised, using a combination of machine learning metrics (e.g., cross-validation, test sets) and formative user testing. While the

machine learning metrics offered useful indicators of directional improvement, they did not capture qualitative shifts in user perceptions.

Formative testing interviews indicated better user experiences after optimizing the NPCEditor cutoff threshold based on real user inputs (e.g., when to fall-through), though these reports were potentially confounded by other advances to system reliability such as more paraphrase examples (Breck et al., 2018). Initially, tests on paraphrases for the first 250 questions recorded indicated a potential advantage for the Semantic w/NN classifier: F-1 statistics (harmonic mean of accuracy and precision) from leave-one-out testing showed the NPCEditor with no off-topic threshold (and no off-topic test examples) showed a substantially lower score (.392) than the Semantic w/NN (.471; Nye et al., 2017). However, in user testing, NPCEditor with a realistic threshold qualitatively outperformed this first Semantic w/NN model in selecting reasonable responses rather than hard misses (entirely irrelevant responses). Ultimately, the threshold was calibrated based on its ability to classify a limited set of real user questions (e.g., by accuracy), then further adjusted based on internal testing (e.g., by perceived quality, including hard misses).

To address RAM limitations, a systematic analysis was also conducted to reduce the Word2Vec corpus size from a full Google News set (about 3.7 GB). Reducing to a very minimal set (89 MB) shifted a classifier based on the top-ranked answer by cosine similarity from 37% to about 29% on a set of 387 holdout paraphrases (Kaimakis et al., 2018). However, the qualitative difference in model performance was only minor for models up to about 300–400 MB (33% accuracy) but increasingly unreliable for models less than that level (despite only showing a drop to 30% accuracy by 100 MB). Based on this analysis, a smaller 300-Slim version was used due to tablet RAM constraints (see: [github.com/eyaler/word2vec-slim](https://github.com/eyaler/word2vec-slim)).

Since providing an “ideal answer” was often less important than avoiding a “hard miss” (obviously irrelevant answer), a random set of responses which were then marked for all questions as either ideal (best possible response) or reasonable (some of the response aligns to the question). Measuring the Semantic w/NN model against this benchmark offered 5-fold cross-validation accuracy scores in the range of 58–67%. However, this underestimates the perceived quality to users, since human ratings for reasonable answers on a set of 218 real user questions at a mentoring center indicated that 80% were reasonable and 73% were ideal and that 110 user-typed questions had an accuracy of 75% reasonable and 67% ideal. While traditional classifier statistics were used to check for the direction of increases or decreases for model quality, this was verified against subjective quality assurance testing.

Overall, direct hit matches and even reasonable matches are not an ideal measure of dialog engine performance in this context, since a large number of mentor answers offer acceptable and reasonable responses to a question (e.g., “What do you like?” could be answered reasonably by either “What do you like about your job?” or “What are your hobbies?” but the first offers more career information). On the other hand, users were often fully content with a “miss” if the mentor’s response was engaging and useful. This strong influence of base-rate relevance for answers to the mentors’ goals implies that future research should consider a scoring function that considers both classification confidence and weightings of importance/relevance.

## Methods

To investigate the impact of this system on student outcomes, a usability study was conducted. The high-level structure of this study was a brief intervention with pre-post surveys for career attitudes and post-only survey measures for usability.

### **Study population**

Participants in this study were high school students attending a summer internship program hosted by a military university. A sample of 31 high school students attending STEM internships participated across three sessions. No interns declined or were unavailable to participate, so any selection bias was solely due to site-selection. Demographic information was not surveyed directly, but participants were able to self-report demographic factors as they related to their desired interest in mentors. The sample composition was majority male (approximately 70% vs. 30% female) and white (approximately 70%, with the remainder primarily Asian or Hispanic).

Participating students already had some interest in STEM, with multiple weeks in the STEM internship prior to the study. This population was less diverse compared to earlier formative testing with MentorPal (e.g., at career fairs and an after-school enrichment program; Breck et al., 2018; Nye et al., 2017). These factors were anticipated to potentially damp changes in career attitudes due to higher career maturity for their age. However, usability measures were expected to be valid, since no qualitative demographic differences were observed during earlier informal testing.

### **Usability study design**

The design was a single condition pre-post. Recruitment was performed by coordinating with the STEM internship program and offering the opportunity to participate. Students were able to opt-out and continue with their normal activities. Participating students were briefed of the project concept and set up on a tablet with a keyboard and a headset with microphone. The students completed a survey, interacted with MentorPanel for 25–30 min, and then completed a post-survey. Students were informed of the three input modes (including voice input). Although students were encouraged to use the system independently some students still interacted occasionally. Researchers unobtrusively observed the students during usage and were available to help if needed. Career confidence and interest measures were evaluated in both the pre and post survey (with alternate item forms) while other items were collected post-survey only.

### **Measures and outcomes**

Study measures were collected through the pre-post survey methodology. Log files were also collected from the users about questions that they asked, which could be aligned to their subject ID. Due to the relatively small sample size and range of questions asked, log file data has not yet shown clear trends to report and will not be discussed in detail in the results section.

Two research questions are derived from Unified Theory of Acceptance and Use of Technology constructs (UTAUT; Venkatesh et al., 2003):

(Q1) Expected Benefit: Will students report positive expected benefit (knowledge, confidence) from virtual mentors?

(Q2) Ease of Use: Will students report positive ease-of-use when interacting with virtual mentors?

These align to UTAUT Performance Expectancy (e.g., that the system will help them do their work better) and Effort Expectancy (e.g., difficulty to use), which influence intent to use a system. Variants of these item prompts have been adapted for learning technology, where Learning Expectancy is a specialized adaptation of Performance Expectancy (e.g., Chen, 2011). To minimize survey fatigue, items for each construct were reduced. Prior research on other virtual agents using this inventory (e.g., Core et al., 2016) used only items with high factor loading from Venkatesh et al. (2003), which showed sufficient consistency and so that subset was adapted here. Career

confidence and knowledge are distinct, so Performance Expectancy was surveyed with two parallel pairs of items which align to Learning Expectancy and Confidence Expectancy. We consider Q1 supported if the average for both Learning Expectancy and Confidence Expectancy are positive (4 or higher on a 6-point scale). Effort Expectancy was likewise reduced to a pair of items that showed strong agreement on Ease of Use. We consider Q2 supported if the average of ease of use items from Effort Expectancy are positive (4 or higher on a 6-point scale). Finally, one item aligned to Attitude (i.e., “is a good idea”) and two for Intent to Use were retained for overall impressions.

The second set of survey items was focused on Mentor-Specific Interest. These items were specific to the mentors, with a pre-post survey about interest in the careers that each mentor represented. In the post-survey only, interest in speaking with each specific mentor as an individual was also collected. Also related to this factor was the list derived from a post-test open response item “If you could talk to a mentor in any career field, which fields would that be?” These were used to determine the degree to which a panel of 4 mentors is sufficient to meet the needs of a cohort of students.

(Q3) Mentor Coverage: Will a panel of four virtual mentors be sufficient to identify at least one helpful mentor to interview and gain knowledge about careers?

We consider Q3 supported if at least 90% of students rate at least one mentor as Helpful or Very Helpful (4 or higher on a 5-point scale). It will also be considered supported if a majority of students find a least one mentor who matches their preferred career fields from that open response field.

Data was also collected to study secondary research areas, specifically the impact of this brief self-guided virtual mentor intervention for (a) changes in their interest in specific careers in the panel and (b) changes in their attitudes about categories of STEM careers. Due to the study sample size and limited number of virtual mentors, these findings are formative. Changes in interest for specific careers were surveyed directly. For broader STEM attitudes, confidence and interest in STEM careers were assessed using variants of the CAPA Career Confidence Inventory (Betz & Borgen, 2010) and the CAPA Interest Inventory, respectively (Betz & Borgen, 2010), which builds on Holland’s (1973) RIASEC set: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. It also measures attitudes toward Science, Technology, Engagement (Individual vs. Team), and related constructs. In consultation with the inventory developer, we created a reduced set of 50 items based on the approximately 400 CAPA items. First, we removed constructs not aligned to any STEM careers in the panel. Second, factor loading data was used to select high-reliability items for the remaining constructs. While the reliability of this item subset has not been measured, this reduced survey should theoretically offer similar inferences. Half of the items aligned to each career area was used for pre and post, to consider attitude changes.

## Results

Usability results adapted from UTAUT are shown in Table 4. Average user ratings were positive but not emphatically positive (“Slightly Agree” or better, but not centered on consistent “Agree”). The strongest ratings were that MentorPal was a good idea, easy to use, and would improve their knowledge. The weakest ratings were that it would change their confidence or interest. Students’ Expected Benefit was positive for Confidence (4.0) and Learning (4.4), which supports the research question that users find this system useful (Q1). Ease of Use was also rated well (4.6), supporting Q2. Intent to Use correlated with Expected Benefit (Pearson  $r = .70$ ,  $p < .001$ ) and Ease of Use (Pearson  $r = .55$ ,  $p < .01$ ).



**Table 4.** Efficacy study post-survey usability mean and standard deviation (scale: 1-strongly disagree to 6-strongly agree).

Usability Survey Construct/Question ( <i>N</i> = 31)	<i>M</i>	<i>SD</i>
<i>Ease of Use</i> ("I found MentorPal easy to use." "Interacting with MentorPal was clear and easy to understand.")	4.6	0.8
<i>Learning Expectancy</i> ("Using MentorPal will help me learn about careers more quickly." "I think MentorPal will improve my knowledge about careers.")	4.4	0.8
<i>Confidence Expectancy</i> ("Using MentorPal will increase my confidence in careers." "MentorPal will help me be more confident about careers.")	4.0	1.0
<i>Attitude Overall</i> ("Using MentorPal is a good idea.")	4.8	1.1
<i>Increased Interest</i> ("MentorPal made me more interested in certain careers.")	3.7	1.2
<i>Intent to Use</i> ("I would use MentorPal if it was part of a course" "I would use MentorPal while looking for colleges or careers")	4.2	1.0

**Table 5.** Frequency of design elements noted in open-response feedback.

	What did you like most?	What would you change?	Difference (Like - Change)
Mentor Design (Video Panel)	45%	6%	+39%
Question Set	39%	23%	+16%
Mentor Diversity	13%	10%	+3%
User Interface Quality	19%	26%	7%
Mentor Careers Represented	6%	16%	10%
Audio/Video Quality	0%	10%	10%
Answer Quality/Classification	19%	32%	13%

Open response feedback offers some context about the elements that users found most valuable versus what they would change (Table 5). The approach of using video mentors in a panel was praised ("actual humans talking, which completely blew me away," "fluid," "less nervous about asking the wrong questions," "time to think about ... what to ask," "differing opinions on the questions"). Students found the question set robust, despite asking for more topics. Opinions of panel diversity were mixed: positive comments were general ("appreciated having mentors of varying ethnicity, gender, and age"), while additional mentors were requested by groups that were not represented (Asians) or under-represented (women). By comparison, few students praised the career coverage, while a larger number said they needed different STEM careers represented. UI feedback was mixed but constructive, even when suggesting changes. Answer quality and classification feedback indicated that despite substantial effort to mitigate this limitation, incorrect answers reduced engagement for some users ("bugs make me less likely to want to use it"). Finally, while most users accepted the low audio-visual production values, they distracted a subset of users.

Users also ranked their preferred modality of input (1 = highest). A two-tailed *t*-test indicated that text input (*M* = 1.46; *SD* = .51) was significantly preferred to voice input (*M* = 2.86; *SD* = .36; *t*(48) 3.35, *p* < .01), but no significant difference from using the button input (*M* = 1.68; *SD* = .72). Prior to the study, this response pattern would have been unexpected. However, observations indicated only one of three sessions used voice input regularly, which may be due to arriving after lunch with more energy. As such, there may be a tipping point where a few users asking questions with their headsets led to other users using speech input. Users may not want to be the first one to start talking in a study group, or they may be concerned about being judged by their peers if they ask questions out loud. Among users who did use the speech input, users found it interesting but were bothered by bugs in the commercial-off-the-shelf speech recognition as compared to the reliability of using typed input or buttons.

**Table 6.** Interest rating for specific mentors (Post-Survey,  $N = 31$ , Scale from 1 = Not at All Interested to 6 = Highly Interested).

Mentor	Career(s)	Interest ratings		% Top rated	% Students listing career in free text
		<i>M</i>	<i>SD</i>		
Clint	CS, Electrical	4.48	1.09	39	23 (13% CS, 13% EE)
Dan	Cryptology, Supercomputers	4.55	1.18	48	6 (3% Cryptology, 3% Supercomputers)
Carlos	Logistics	2.65	1.56	10	0
Julianne	Systems Engineering, Aviator	3.81	1.47	23	6 (Aviation)

### ***Mentor-specific interest***

The level of interest each student reported for mentors after using MentorPal is noted in Table 6. Mentors in computer-related fields were rated highest, while logistics was rated the lowest. These patterns are reproduced when considering a student's top-rated mentor, who must be Helpful or Very Helpful (4 or higher out of 5), except that even lesser-rated mentors still represent the favorite mentor for a subset of students. Overall, 87% of students rated at least one mentor Helpful or better. This was not quite enough to cross the threshold for Q3, which targeted 90% of students finding a good match in a panel of 4 mentors. However, student open responses give good indicators on how to increase the likelihood of a good virtual mentor match.

On the right side of Table 6, the mentor's expertise is compared to match to user's answers to "If you could talk to a mentor in any career field, which fields would that be?" Most students answered with multiple career fields (26% one field, 26% two fields, 45% three or more), so this column is not additive. Overall, 29% of student free text responses identified listed at least one career available in the panel. The three categories preferred but not included were: Psychology/Cognitive Science (10%), Biology/Medicine (19%), and Finance/Business (10%). Had such mentors been available, then 55% of students would have had at least one mentor in their preferred fields. Within a panel of any four specific mentors, the optimal panel of 4 mentors would have still only match 48% of students with a preferred field, which indicates the need for students to interact with a personalized set of mentors.

### ***STEM career attitudes: Changes in interest and confidence***

As an exploratory analysis for potential impact, Table 7 presents the change in interest in STEM fields from the virtual mentoring session. In this context, two-tailed tests were employed because either increases or decreases may represent increased career maturity, since students may be focusing their interests better. Non-significant increases were observed for Supercomputers, Military Careers, and STEM fields overall. A significant drop in interest was observed for Logistics. This decrease aligns reasonably with the students' existing career goals, which were fairly heavily concentrated on topics such as biomedicine, engineering, and computer science.

Pre-post difference in career attitudes for different types of careers is shown in Table 8. These aggregates focus on the career types and most contain one interest item and one confidence item per survey. Unfortunately, when analyzing the item results, it was found that certain items targeting the same construct on the pre-survey in some cases showed substantially different base rates for this population (e.g.,  $>0.5$  average difference). This raises questions about the validity of comparing pre-post differences, since the differences in base rates between items might be comparable to the differences resulting from the outcomes. This issue can only be resolved by conducting future studies with the full attitude survey for both pre and post, or collecting a larger data set where the two smaller versions are randomized such that the base rates for each item can be determined. For constructs where the items can be assumed to reliably capture the same

construct, it would imply that the MentorPanel may have increased individual engagement but decreased enterprising attitudes (e.g., sales, management).

**Table 7.** Change in interest in STEM and the career fields represented ( $N = 31$ ).

Career(s)	Pre-survey interest ratings		Post-survey interest ratings		Difference
	M	SD	M	SD	
Computer Science	3.5	1.2	3.55	1.36	0.05
Electrical Engineering	2.97	1.25	3.03	1.27	0.06
Logistics	3.13	0.81	2.71	0.97	0.42
Supercomputers/High Performance Computing	3.03	1.43	3.19	1.33	0.16
Systems Engineering	2.94	1.00	2.87	0.92	0.07
STEM Overall	4.35	0.88	4.48	0.85	0.13
Military	2.55	1.46	2.74	1.41	0.19

$p < 0.05$ .  
 $p < 0.01$  on two-tailed  $t$ -test.

**Table 8.** Change in interest and confidence about certain career traits (from 1 = No Confidence/Strongly Dislike to 5 = Complete Confidence/Strongly Like;  $N = 31$ ).

Career Type	Pre-survey interest & confidence		Post-survey interest & confidence		Difference
	M	SD	M	SD	
A - Artistic	3.1	0.7	3.2	0.8	0.2
C - Conventional	3.1	0.6	3.1	0.8	0.0
E - Engineering	3.4	1.0	3.0	0.9	0.4
Individual Engagement	3.4	0.6	4.1	0.6	0.7
Opp I - Enterprising	3.6	0.7	2.6	0.8	0.9
Opposite R - Social	3.8	0.6	3.5	0.7	0.3
R - More	3.7	0.7	3.7	1.1	0.0
S - Science	3.5	0.7	3.5	0.9	0.0
T - Technology	3.3	1.2	3.0	1.1	0.3
Team Engagement	4.0	0.5	3.7	0.6	0.3

$p < .05$ .  
 $p < .01$  on two-tailed  $t$ -test. See the associated narrative for caveats in interpreting the above table, due to issues in item reliability as they were used here.

**Discussion**

This study indicates that users found MentorPal to have a positive perceived benefit for both their career knowledge and confidence (Q1) and that the system had an acceptable ease-of-use (Q2). Their positive attitudes persisted despite a dialog classifier with limited accuracy (only 75% of user-input questions classified with a reasonable answer) and production quality on-par with a video conference. This user acceptance of relatively raw virtual mentors is critical because MentorPal is researching techniques for a rapid, turnkey mentor pipeline that would enable a large number of virtual mentors. The importance of such a pipeline was also obvious in student comments, where greater diversity in both careers and demographic representation were key requests.

A notable finding for usability in a virtual panel was that even a partially-recorded member, with a small number of questions, can be effective in a panel. For example, due to external events the “Julianne” mentor had only a limited subset of questions. While some students that mentor unsatisfying due its limited response repetoire, the majority of students either did not notice a difference or even found Julianne to be their favorite mentor. This was particularly evident because some of the female high school students strongly preferred interacting with female mentors. Discovering this ability to integrate mentors with limited responses is an important finding, because the time to record a full mentor is substantial (12–20h) and it impacts what mentors

were available and willing. However, if mentors with limited content can be integrated with more robust mentors, panels could be designed more efficiently by pairing larger general-purpose mentors with mentors with specialized experiences.

The research question (Q3) that a panel of four mentors would result in a large majority (e.g., at least 90%) of students finding a well-matched mentor was not conclusive. This specific panel only found satisfactory matches for 87% of students and ideal preferences (from open response) for 29% of participants. Moreover, no panel of with four careers could match preferred mentors for even half of students. This demonstrates the importance of a scalable process that would allow students to select a personalized panel drawn from a larger set of recorded mentors. This issue is particularly salient when considering diversity and representation issues, where the range of characteristics implies a combinatorial explosion (e.g., gender, sexual orientation, race, age, geography, attitudes/beliefs). With that said, students may not require one mentor with all their ideal criteria, but that they might still benefit substantially from a mentor in their preferred STEM career and another mentor with a similar demographic in an adjacent field.

MentorPal's impact on career attitudes was indeterminate. Small increases were observed for most of the specific careers represented (Supercomputing, STEM Careers, Military Careers), while in one field (Logistics) interest decreased. In general, since choosing a career is a selection task, greater career maturity should simultaneously increase interest in some careers but decrease interest in the traits and fields that one should not pursue. However, it is unclear what set of increases and decreases represent progress toward a *better* career interest (for that student) without longer-term outcome data. Second, as a confound, changes might be either due to increased career insight or due to mentor-specific perceptions. For example, students appeared unfamiliar with Logistics and may have found that the field did not fit into their pre-conceived STEM interests. However, it could be due to mentor-specific factors such as voice (disliked by one student) or his description of logistics from a manager's role (e.g., leading a team, coordinating with sponsors, and supply chain timing). Results from the CAPA inventory had only a small sample and were similarly inconclusive, in that while some significant shifts occurred (e.g., Individual Engagement, Enterprising), the importance of attitude changes may depend on the individual (e.g., a student just learning about STEM fields vs. one who is already preparing for a specific field).

### **Contributions: Virtual mentor agents**

This work contributes most strongly to research on virtual mentor agents, both in showing a novel application for how they may be used, and a new technical approach to creating them that lowers the cost and effort required but still creates useful agents. Specific contributions are made on the following topics.

*Multi-Agent Pedagogy:* The MentorPanel design contributes to the growing topic of multi-agent pedagogical systems (Kim & Baylor, 2016), such as AutoTutor dialogues (Cai et al., 2011). Participants appreciated that the panel was “non-judgmental,” offered “different viewpoints,” and “all aspects of life” for a career. To our knowledge, this is the first pedagogical agent system to train multiple agents for an ensemble panel.

*Cold-Start Mitigation:* A common problem for dialog systems is poor initial performance due to lack of user data (Yan et al., 2017). When rapidly developing mentors, each MentorPal mentor suffers a partial new-system cold start and mediocre classification accuracy for user questions ( $\leq 75\%$ ), which was a significant step down from other video-based virtual humans (e.g., Artstein et al., 2014). However, four strategies helped users accept the system: a more-constrained domain (careers vs. entire-life), topic buttons to cue learners on relevant questions and as a reliable fallback, a shared set of general paraphrases for questions used by all mentors, and combining individually trained mentors as an ensemble (so less confident can remain silent). While the

gain from each strategy has not been isolated, these strategies can be re-used and studied in future systems.

*Generalist vs. Specialist Agent Role:* Another finding from this research was that in a panel setting, users accepted agents with a more limited question set even though they could detect the difference. This implies a pedagogical agent role that may be a subset of “Expert” (Baylor & Kim, 2016): generalist vs. specialist. To study this, the next version of MentorPal offers a “Recommended” question set for a conversation, which may enable specialists to contribute more effectively by focusing the conversation domain.

*Mentor Career Coverage:* This research also studied how well a panel was able to cover the career interests of a specialized population (e.g., high school students with preexisting STEM career interest). In this sample, which is more likely to have specific interests (26% only interested in one mentor career), even the best hypothetical set of careers had an upper bound of less than 50% coverage, even where some mentors covered multiple careers. As such, this means that personalized panels will be required to adequately cover students’ career interests and demographic representation. Finally, for this population, student comments focused more on having their preferred careers represented rather than shallow similarity factors (in alignment with Eby et al., 2013). However, a few students offered clear open-response feedback when their own salient identity was not represented or substantially under-represented. This implies that a well-balanced personalized panel might contain at least two mentors from preferred fields and also up to two mentors who ensure representation of underrepresented students’ identities.

## Limitations

Generalizing findings from this work are limited by the study design and complexity of mentoring as a pedagogical domain. Specific limitations include the sample size, sample diversity, and lack of evidence of student impact.

*Sample Size:* While the sample size is appropriate for a usability study that demonstrates the feasibility of MentorPal, this sample size is insufficient to assess the impact of mentors on students’ career attitudes or behavior. Specifically, it is expected that mentors serve different implicit roles for students based on their initial career maturity: undecided students may need to explore career options (shift attitudes and knowledge on a broad set of careers), while students with a clear STEM career goal may instead need strategic advice (focus on one specific career). Larger sample sizes would enable studying how these interactions moderate mentoring impact on students and how impact should be evaluated for students at different stages of career development.

*Sample Diversity:* A second limitation of this usability study is that the recruitment site was biased toward students with preexisting STEM engagement and whose demographics align largely to existing STEM pipelines (as opposed to under-represented minorities). This means that findings from this population might not necessarily generalize to under-represented minority groups for STEM. With that said, in less-formal usability tests with a larger and more diverse group at a STEM fair (e.g., Breck et al., 2018), no clear trends were observed for difficulties in usability (Q1). In that work, researchers noted a possible trend where students with clearer career goals (e.g., older) are more specific and pickier about mentor content, but if that were true a broader sample might give higher Perceived Benefit ratings (Q2). The main limitation is that students’ preferred panel coverage (Q3) almost certainly over-represents the importance of specific career coverage vs. demographic representation. This also means that other key characteristics or roles for virtual mentors may have been missed during this study.

*Impact:* Finally, this usability study was not designed to offer conclusive evidence that virtual mentors make a meaningful impact on career decisions. While it is assumed that such mentors should produce outcomes similar to those of a career fair, validating impact would require assessment of student career mentoring needs (e.g., prior knowledge, strength of career goals), salient

identities for students (e.g., shallow and deep similarity), and also longer-term measures (e.g., changes in study effort, class enrollment, placement in STEM majors). Expanded collection of existing inventories (e.g., using all 50 CAPA-based items in both pre and post surveys) would also help identify changes in attitudes with greater reliability.

## Conclusion

Based on the results of this study, it was found that a question set of approximately 400 responses could support an adequate 5–20-min question-and-answer conversations about career fields. While our goal of creating mentors inexpensively and quickly did reduce response accuracy and video production quality as compared to more expensive approaches, high school students still found the experience compelling and useful. Moreover, the MentorPanel showed that larger “generalist” mentors might be complemented by “specialist” mentors (under 100 questions), which might help increase exposure to distinct careers and life experiences for mentors who might be unavailable for multiple days of recording. A final finding was that a panel of four mentors (even one hypothetically optimized through hindsight) is insufficient to cover either the main career interests or diversity representation of even 31 students. This means that future research should investigate how students respond to self-selecting or automatically personalized panels drawn from a larger set of mentors representing broader career choices and backgrounds.

Having established that this approach can lead to mentors which students find usable and beneficial, the MentorPal framework could be used to study a number of broader questions: (a) modeling mentor influence, (b) investigating change mechanisms, and (c) exploring the impact of outreach to under-represented groups. On the first topic, the MentorPanel can be used to study the factors predict student’s engagement with mentors (e.g., interest in a mentor due to their career, personality, or perceived similarity)

While the current study could not distinguish between interest in a career versus a specific mentor, this might be explored with a study containing a balanced set of mentors (e.g., same field, diverse mentors). However, even in that case, multifaceted traits such as deep similarity or mentor personality may overwhelm shallower differences (Eby et al., 2013). As a complementary study, research on mechanisms such as stereotype-threat might study changes in confidence or test outcomes as determined by the number of panelists with similar identity characteristics. For example, for the Asch conformity protocol, it was found that the conformity effect was strongly diminished by one minority vote and nearly eliminated by two (Asch, 1952). This is complicated in a panel, where each mentor may share some but not all identities with a student. This effect could be studied initially with a simpler protocol with only topic buttons, such that each agent might only have 10 questions (to enable recording a large volume of mentors).

Second, this usability study did not investigate mechanisms of change for the impact of mentors. However, based on insights from the study, one productive direction may be to study the impact on student attitudes based on their career maturity. Careers where students indicate the least knowledge or commitment would be expected to shift their attitudes more strongly for that career and would be more likely to browse mentors. Students with preexisting high levels of career maturity would be expected to not change as much for attitudes, but instead to focus on a matching mentor and increased knowledge. These differences should also be evident in the types of questions asked.

This issue impacts how virtual mentors can contribute to a larger pedagogical framework. The usability study presented above describes the results for when students are largely interacting with MentorPal as a free-play simulation. However, research indicates that entirely unstructured simulations tend to produce lower learning gains than when users receive supporting pedagogy such as framing the goals, goal-setting, and reviewing simulation outcomes with feedback (Chang,



2017). Future research must study how a scalable mentor panel or virtual career fair should be integrated into a larger career development process.

Finally, a broader and more diverse study is required to better understand its efficacy, particularly for under-represented minorities in STEM and among students who were not previously considering STEM careers. A core capability for MentorPal is that it should be able to exemplify mentors who are not well represented in a career. As noted in the Limitations section, validating the impact of MentorPal requires a study that combines student needs assessments, more comprehensive immediate impact measures, and follow-up metrics to determine if virtual mentor conversations contribute to their longer-term development of career goals. While the MentorPal virtual agent mentoring presents a promising first step, it must be studied as a tool in the context of larger career development strategies and societal outcomes.

## Acknowledgments

The statements and views in this paper are the views of the authors alone. Additionally, many thanks to the USC Armani Lab who organized the EngX STEM fair where MentorPal was exhibited and to the Naval Postgraduate School program which engaged with us on this research. We would also like to extend our thanks to Fred Borgen's insight into his CAPA inventory items.

## Disclosure statement

The authors declare that they have no conflicts of interest in this work.

## Funding

This work was supported by grants from the Office of Naval Research's STEM Program (ONR N00014-16-1-2820) and by students supported by the National Science Foundation Research Experience for Undergraduates program (NSF 1560426).

## Notes on contributors

**Benjamin Nye**, Ph.D. is the Director of Learning Science at the University of Southern California, Institute of Creative Technologies. Ben's research tries to remove barriers to development and adoption of adaptive and interactive learning technology so that they can reach larger numbers of learners. Dr. Nye's research has been recognized for excellence in intelligent tutoring systems, cognitive agents, and realistic behavior in training simulations. His research is on scalable learning technologies and design principles that promote learning. He is the membership chair for the International of Artificial Intelligence in Education (IAIED) Society and holds memberships in Educational Data Mining Society (EDM), and Association for the Advancement of Artificial Intelligence (AAAI).

**Dan M. Davis** is a consultant for USC ICT, focusing on large-scale distributed DoD simulations. At USC's Information Sciences Institute, he was the Director of the JESPP project for JFCOM for a decade. As the Assistant Director of the Center for Advanced Computing Research at Caltech, he managed Synthetic Forces Express, bringing HPC to DoD simulations. Prior experience includes serving as a Director at the Maui High Performance Computing Center and as a Software Engineer at the Jet Propulsion Laboratory and Martin Marietta. He has served as the Chairman of the Coalition of Academic Supercomputing Centers and has taught at the undergraduate and graduate levels. As early as 1971, Dan was writing programs in FORTRAN on one of Seymour Cray's CDC 6500's. He saw duty in Vietnam as a USMC Cryptologist and retired as a Commander, Cryptologic Specialty, U.S.N.R. He received B.A. and J.D. degrees from the University of Colorado in Boulder.

**Sanad Z. Rizvi** was a visiting Research Assistant at the Institute for Creative Technologies of the University of Southern California. His interests are in Machine Learning Research and its applications in the real world. Given his Computer Science background, he has accumulated good experience in software development through various research activities and programming projects. His expertise in Python has allowed him to make open-source contributions to that language. Currently, he is focusing on Deep Learning research for Natural Language Processing and is an Instructor for India's prominent MOOC for Natural Language Processing. He earned a Bachelor of

Technology degree in Computer Science and Engineering from The National Institute of Engineering, Mysore, India.

**Kayla Carr** is a software analyst for USC ICT. She has worked on PAL3 and MentorPAL, building mobile and web based learning applications for sailors and other learners. She graduated with a Masters in Computer Science from the University of Southern California and started at ICT as a student researcher during her degree.

**William Swartout**, Ph.D. is Chief Technology Officer and co-founder of the USC Institute for Creative Technologies and a research professor in the Computer Science Department at the USC Viterbi School of Engineering. His research interests include virtual humans, explanation and text generation, knowledge acquisition, knowledge representation, intelligent computer-based education and the development of new AI architectures. In 2009, Swartout received the Robert Englemore Award from the Association for the Advancement of Artificial Intelligence (AAAI). Swartout is a Fellow of the AAAI, has served on their Board of Councilors, and is past chair of the Special Interest Group on Artificial Intelligence (SIGART) of the Association for Computing Machinery (ACM).

**Raj Thaker** is in the MS in Computer Science program at the University of Southern California. His focus is centered around backend infrastructure development with primary contributions in Java frameworks. He worked with Amazon AWS Internet of Things (IoT) during Summer 2019 working on developing an end-to-end pipeline for data forecasting. During his time at USC, Raj collaborated on research projects with the Institute for Creative Technologies on open-source intelligent tutoring system prototypes. Apart from personal and commercial projects, Raj has worked to provide technology solutions to NGOs like Happy Hearts Foundation and 5050 Leadership and has been a technical judge at coding competitions like Code for Good by JPMorgan Chase.

**Kenneth Shaw** was a visiting Research Assistant at the Institute for Creative Technologies of the University of Southern California, selected under the National Science Foundation's Research Experiences for Undergraduates in summer 2018, where he worked on starting a web-enabled version of MentorPal. His current research includes working on safe, efficient human-robot collaboration at Carnegie Mellon University and heterogeneous multi-agent robotics task allocation at Georgia Tech under the Army Research Lab DCIST project. He is currently in his final year studying Computer Engineering and Computer Science in Intelligence at the Georgia Institute of Technology.

## ORCID

Benjamin D. Nye  <http://orcid.org/0000-0002-5902-9196>

## References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P. ... Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation* 16 (pp. 265–283).
- Amanatullah, E. T., & Morris, M. W. (2010). Negotiating gender roles: Gender differences in assertive negotiating are mediated by women's fear of backlash and attenuated when negotiating on behalf of others. *Journal of Personality and Social Psychology*, 98(2), 256–267. <https://doi.org/10.1037/a0017094>
- Arroyo, I., Woolf, B. P., Cooper, D. G., Burleson, W., & Muldner, K. (2011). The impact of animated pedagogical agents on girls' and boys' emotions, attitudes, behaviors and learning [Paper presentation]. In *2011 IEEE 11th International Conference on Advanced Learning Technologies* (pp. 506–510). IEEE.
- Artstein, R., Traum, D., Alexander, O., Leuski, A., Jones, A., Georgila, K., Debevec, P., Swartout, W. R., Maio, H., & Smith, S. (2014). Time-offset interaction with a Holocaust survivor [Paper presentation]. In *ACM Conference on Intelligent User Interfaces (IUI) 2014* (pp. 163–168). ACM Press. <https://doi.org/10.1145/2557500.2557540>
- Asch, S. E. (1952). *Social psychology*. Prentice-Hall.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.
- Baylor, A. L. (2009). Promoting motivation with virtual agents and avatars: Role of visual presence and appearance. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364(1535), 3559–3565. <https://doi.org/10.1098/rstb.2009.0148>
- Baylor, A. L., & Kim, Y. (2016). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, 26(1), 160–115. <https://doi.org/10.1007/s40593-015-0055-y>
- Breck, S., Carr, K., Davis, D. M., Nordhagen, J. N., & Nye, B. D. (2018). Virtual Mentors in a Real STEM Fair: Experiences, Challenges, and Opportunities [Paper presentation]. In *International Conference on Artificial*

- Intelligence in Education (AIED) 2018: Workshop on Intelligent Mentoring Systems (IMS 2018) Proceedings* (pp. 1–12).
- Betz, N. E., & Borgen, F. H. (2010). The CAPA integrative online system for college major exploration. *Journal of Career Assessment*, 18(4), 317–327. <https://doi.org/10.1177/1069072710374492>
- Bickmore, T., Schulman, D., & Yin, L. (2010). Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence: Aai*, 24(6), 648–666. <https://doi.org/10.1080/08839514.2010.492259>
- Bieber, A. E., Marchese, P., & Engelberg, D. (2005). The laser academy: An after-school program to promote interest in technology careers. *Journal of Science Education and Technology*, 14(1), 135–142. <https://doi.org/10.1007/s10956-005-2741-2>
- Cai, Z., Graesser, A. C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., Halpern, D., & Butler, H. (2011). *Trialog in ARIES: User input assessment in an intelligent tutoring system*. In Proceedings of the 3rd IEEE international conference on intelligent computing and intelligent systems (pp. 429–433). IEEE Press.
- Chang, H. Y. (2017). How to augment the learning impact of computer simulations? The designs and effects of interactivity and scaffolding. *Interactive Learning Environments*, 25(8), 1083–1097. <https://doi.org/10.1080/10494820.2016.1250222>
- Chen, J. L. (2011). The effects of education compatibility and technological expectancy on e-learning acceptance. *Computers & Education*, 57(2), 1501–1511. <https://doi.org/10.1016/j.compedu.2011.02.009>
- Core, M. G., Georgila, K., Nye, B. D., Auerbach, D., Liu, Z. F., & DiNinni, R. (2016). Learning, adaptive support, student traits, and engagement in scenario-based learning [Paper presentation]. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2016*. Paper 16288 (pp. 1–13). Springer.
- Dahlstrom, M. F. (2014). Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences of the United States of America*, 111(Supplement\_4), 13614–13620. <https://doi.org/10.1073/pnas.1320645111>
- Eby, L. T. d T., Allen, T. D., Hoffman, B. J., Baranik, L. E., Sauer, J. B., Baldwin, S., Morrison, M. A., Kinkade, K. M., Maher, C. P., Curtis, S., & Evans, S. C. (2013). An interdisciplinary meta-analysis of the potential antecedents, correlates, and consequences of protégé perceptions of mentoring. *Psychological Bulletin*, 139(2), 441–476. <https://doi.org/10.1037/a0029279>
- Fries-Britt, S., & Snider, J. (2015). Mentoring outside the line: The importance of authenticity, transparency, and vulnerability in effective mentoring relationships. *New Directions for Higher Education*, 2015(171), 3–11. <https://doi.org/10.1002/he.20137>
- Herman, S. (2010). Career HOPES: An Internet-delivered career development intervention. *Computers in Human Behavior*, 26(3), 339–344. <https://doi.org/10.1016/j.chb.2009.11.003>
- Graesser, A. C., Swamer, S. S., Baggett, W. B., & Sell, M. A. (2014). New models of deep comprehension. In B. K. Briton, & A. C. Graesser (Eds.), *Models of understanding text* (pp. 9–40). Psychology Press.
- Gratch, J., Lucas, G. M., King, A. A., & Morency, L. P. (2014). It's only a computer: The impact of human-agent interaction in clinical interviews [Paper presentation]. In *Autonomous Agents and Multi-Agent Systems (AAMAS) 2014 Conference* (pp. 85–92). International Foundation for Autonomous Agents and Multiagent Systems.
- Hampton, A. J., Nye, B. D., Pavlik, P. I., Swartout, W. R., Graesser, A. C., & Gunderson, J. (2018). *Mitigating knowledge decay from instruction with voluntary use of an adaptive learning system*. In International Conference on Artificial Intelligence in Education (pp. 119–133). Springer.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735> PMID 9377276.
- Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Prentice-Hall.
- 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(1), 25–36. <https://doi.org/10.1007/s40593-015-0065-9>
- Kaimakis, N. J., Davis, D. M., Breck, S., Nye, B. D. (2018). Domain-specific reduction of language model databases: Overcoming Chatbot implementation obstacles. In *The Proceedings of the ModSim World Conference*.
- Kim, Y., & Baylor, A. L. (2016). Research-based design of pedagogical agent roles: A review, progress, and recommendations. *International Journal of Artificial Intelligence in Education*, 26(1), 160–169. <https://doi.org/10.1007/s40593-015-0055-y>
- Kraiger, K., Finkelstein, L. M., & Varghese, L. S. (2019). Enacting effective mentoring behaviors: Development and initial investigation of the cuboid of mentoring. *Journal of Business and Psychology*, 34(4), 403–424. <https://doi.org/10.1007/s10869-018-9551-z>
- Krieshok, T. S. (1998). An anti-introspectivist view of career decision making. *The Career Development Quarterly*, 46(3), 210–229. <https://doi.org/10.1002/j.2161-0045.1998.tb00697.x>
- Lehming, R., Gawalt, J., Cohen, S., & Bell, R. (2013). *Women, minorities, and persons with disabilities in science and engineering: 2013* (NSF Special Report 13-304). NSF.

- Leuski, A., & Traum, D. (2011). NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2), 42–56. <https://doi.org/10.1609/aimag.v32i2.2347>
- Macdonald, A. (2014). “Not for people like me?” Under-represented groups in science, technology and engineering. Wise Campaign.
- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19(4–5), 494–513. <https://doi.org/10.1080/09541440701326154>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.
- Noonan, R. (2017). *STEM jobs: 2017 update* (ESA Issue Brief# 02-17). US Department of Commerce.
- Nye, B., Swartout, W., Campbell, J., Krishnamachari, M., Kaimakis, N., Davis, D. (2017). MentorPal: Interactive virtual mentors based on real-life STEM professionals. In *The Proceedings of the Interservice/Industry Simulation, Training and Education Conference*, 2017.
- Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research*, 49(1), 1–39. <https://doi.org/10.2190/EC.49.1.a>
- Stoeger, H., Duan, X., Schirner, S., Greindl, T., & Ziegler, A. (2013). The effectiveness of a one-year online mentoring program for girls in STEM. *Computers & Education*, 69, 408–418. <https://doi.org/10.1016/j.compedu.2013.07.032>
- Swartout, B., Nye, B. D., Hartholt, A., Reilly, A., Graesser, A. C., VanLehn, K., Wetzel, J., Liewer, M., Morbini, F., Morgan, B., Wang, L., Benn, G., & Rosenberg, M. (2016). Designing a personal assistant for life-long learning (PAL3) [Paper presentation]. In *Florida Artificial Intelligence Research Society (FLAIRS) 2016 Conference* (pp. 491–496). AAAI Press.
- Swartout, W., Artstein, R., Forbell, E., Foutz, S., Lane, H. C., Lange, B., Morie, J. F., Rizzo, A. S., & Traum, D. (2013). Virtual humans for learning. *AI Magazine*, 34(4), 13–30. <https://doi.org/10.1609/aimag.v34i4.2487>
- Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J.-Y., Gerten, J., Chu, S., & White, K. (2010). Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides [Paper presentation]. In *Intelligent Virtual Agents (IVA) 2010 Conference* (pp. 286–300). Springer.
- Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D. & Swartout, W. (2012). Ada and Grace: Direct interaction with museum visitors [Paper presentation]. In *Intelligent Virtual Agents (IVA) 2012 Conference* (pp. 245–251). Springer.
- Traum, D., Jones, A., Hays, K., Maio, H., Alexander, O., Artstein, R., Debevec, P., Gainer, A., Georgila, K., Haase, K., Jungblut, K., Leuski, A., Smith, S. & Haase, K. (2015). New dimensions in testimony: Digitally preserving a Holocaust Survivor’s interactive storytelling. In *Interactive Storytelling* (pp. 269–281). Springer.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27, 425–478. <https://doi.org/10.2307/30036540>
- Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, 50(5), 1081–1121. <https://doi.org/10.3102/0002831213488622>
- Wellener, P., Dollar, B., Manolian, H. A. (2018). The future of work in manufacturing. *Deloitte Insights*. [www2.deloitte.com/insights/us/en/industry/manufacturing/future-of-work-manufacturing-jobs-in-digital-era.html](http://www2.deloitte.com/insights/us/en/industry/manufacturing/future-of-work-manufacturing-jobs-in-digital-era.html)
- Yan, Z., Duan, N., Chen, P., Zhou, M., Zhou, J., Li, Z. (2017). Building task-oriented dialogue systems for online shopping. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)* (pp. 4618–4625). AAAI Press.