MentorPal: Interactive Virtual Mentors Based on Real-Life STEM Professionals

Benjamin D. Nye, Nicholas J. Kaimakis,
Madhusudhan Krishnamachari, William Swartout,
Julia Campbell, and Clinton Anderson
Institute for Creative Technologies, University of Southern California
Los Angeles, California
{nye, kaimakis, mkrishnamachari, swartout, campbell, canderson}@ict.usc.edu

Dan M. Davis
High Performance Computing Education
Long Beach, California
dmdavis@acm.org

ABSTRACT

In an ideal world, all students could meet STEM role models as they explore different careers. However, events such as career fairs do not scale well: professionals have limited time and effective mentors are not readily available in all fields. The result is that students' understanding is minimal about what professionals in STEM fields do every day, what education is needed, and even what STEM fields exist. Moreover, since in-person interactions rely on finding people engaged in current STEM careers, students may form career goals for stagnant fields rather than growing fields (e.g., projected workforce needs). To address this problem, we are designing a scalable tablet-based app that gives students the opportunity to converse with interactive recordings of real-life STEM professionals. These conversational virtual agents will emulate a question-and-answer session with STEM professionals who have Navy ties and who are engaging, enthusiastic, and effective mentors. These interactions will allow students to have a life-like informational interview with a virtual agent whose responses are directly drawn from a specific real professional's video-recorded interview. This work differs from prior research on career guides by capturing the experiences of a collection of unique mentors, which should be more authentic and engaging than a generic agent or resource which speaks only about the average experience. This paper will discuss the process of creating the first such virtual STEM mentor prototype, including the development of an extensive mentoring question bank (approximately 500 questions); key mentoring topics that intersect STEM, DoD, and civilian life; techniques for cost-effective recording of remote mentors; and the process of training and verifying a natural language dialogue model for answering and suggesting career questions. Finally, we conclude with implications, strengths, and drawbacks of virtualizing the experience of talking with specific mentors, from the perspectives of efficacy, scalability, and maintainability.

ABOUT THE AUTHORS

Benjamin Nye, Ph.D. is the Director of Learning Science at the University of Southern California, Institute of Creative Technologies (USC ICT). 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 (1st Place ONR ITS STEM Grand Challenge), cognitive agents (BRIMS 2012 best paper), and realistic behavior in training simulations (Federal Virtual Worlds Challenge; Silverman et al., 2012). His research is on scalable learning technologies and design principles that promote learning. This research has led to 25 peer-reviewed papers, 11 book chapters, and 5 open-source projects. He is the membership chair for the International Artificial Intelligence in Education (IAIED) Society and holds memberships in Educational Data Mining Society (EDM), and Association for the Advancement of Artificial Intelligence (AAAI). He also co-chairs the FLAIRS Learning Technologies track (2015-2017).

Nicholas J. Kaimakis is an undergraduate student studying Computer Science within the Viterbi School of Engineering at the University of Southern California. His current research thrusts are in the use of computer generated avatars or video clips, animated and directed by natural language optimized Artificial Intelligence (A/I) programs that present a life-like dialogue capability to interact with remote users via the internet. He has designed efficient project structures, taught programming, and managed interdisciplinary teams. His current project is funded by the Navy and is designed to help improve knowledge of STEM fields across varied demographics with the development of an interactive interface that makes STEM information more accessible on-line.

Madhusudhan Krishnamachari is a Research Assistant at USC ICT. His current research interests are in conversational Artificial Intelligence and the development of virtual agents that can engage in dialogue with humans. He has demonstrated success in designing dialogue policies for virtual agents, developing algorithms to answer questions in conversations and in managing the development lifecycle of the dialogue component of virtual agents. He earned his B.Tech. in Computer Science from Amrita University, India and his M.S. in CS at the University of Southern California in May of 2017.

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 Engelmore 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).

Julia C. Campbell, Ed.D. is a Project Director at USC ICT. Her work focuses on cognitive task analysis to improve performance and self-efficacy. More specifically, she has conducted ongoing research aimed at understanding and improving distributed learning through a series of recommendations and best practices for promoting and enhancing student engagement online. She received a M.A. in Communication and Ed.D. (in Education) from USC in 2006 and 2010, respectively.

Clinton Anderson is an Assistant Language Teacher for elementary and junior high school students in northern Japan. He is a supervisor for a 35-member unit as a Chief Electrician's Mate in the US Navy Reserve, and has eight years of active duty experience as a nuclear-trained Electrician on an aircraft carrier. His simulation and modeling interests have allowed him to work on the DoD's Multi-Purpose Supporting Arms Trainer at the Naval Air Warfare Center Training Systems Division and a novel text-input method at Tohoku University. Recently, he has participated in the development of military school virtual education at USC ICT. Clinton holds a B.A. Degree in Computer Science (CS) from UC Berkeley (Education minor) and is in the CS Master's program in Game Development within the Viterbi School of Engineering at USC.

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.

MentorPal:
Interactive Virtual Mentors Based on Real-Life STEM Professionals

Benjamin D. Nye, Nicholas J. Kaimakis,
Madhusudhan Krishnamachari, William Swartout,
Julia Campbell, and Clinton Anderson
Institute for Creative Technologies, University of Southern California
Los Angeles, California
{nye, kaimakis, mkrishnamachari, swartout,
campbell, canderson}@ict.usc.edu

Dan M. Davis
High Performance Computing Education
Long Beach, California
dmdavis@acm.org

INTRODUCTION

Despite efforts by career counselors to provide information to K-12 students about Science, Technology, Engineering, and Mathematics (STEM) careers, students' understanding about what professionals in STEM fields do on a day-to-day basis, what education is needed, and even what STEM fields exist is minimal (Bieber et al., 2005). This prevents many underrepresented students from entering STEM, since students' opportunities are impacted by their science and mathematics achievements as early as high school (Wang, 2013).

In an ideal world, guidance counselors and teachers would be able to easily organize activities such as career fairs, where every student could talk with representative STEM role models. Research has shown that such activities are highly-effective for increasing awareness, motivation, and engagement toward STEM careers (Brown, Ryan Krane, Brecheisen, Castelino, Budisin, Miller, & Edens, 2003). However, they do not scale well: professionals have limited time, and those who volunteer tend to be connected to the school (e.g., parents, friends of staff). As such, the students who need these types of experiences the most will be the least likely to receive them. Moreover, since in-person interactions rely on finding people with current STEM careers, students may form career goals for stagnant fields rather than growing fields that are relevant to the future (e.g., projected workforce needs).

Fundamentally, there is a need for a *role model amplifier*: a technology or methodology that enables students to reproduce the value of a frank, one-on-one discussion with real-life professionals so that students can understand more about what a career really requires, what the work is like, and why that career might (or might not) be a good match for them. To address this problem, we are designing MentorPal: a scalable tablet-based app that gives students the opportunity to interact with interactive recordings of real-life STEM professionals.

These conversational virtual agents will emulate a question-and-answer session with STEM professionals who have Navy ties and who are engaging, enthusiastic, and effective mentors. These interactions will allow students to have a life-like informational interview with a virtual agent whose responses are directly drawn from a specific real professional's video-recorded interview. This work differs from prior research on career guides by capturing the experiences of a collection of unique mentors, which should be more authentic and engaging than a generic agent or resource which speaks only about the average experience.

Research on the MentorPal system is currently in progress, with this first stage focused on the iterative development of an interactive agent for the first mentor, a Chief Electrician's Mate (EMC) currently serving in the Navy Reserve. In the upcoming stages, this technology will be expanded to host a panel of four distinct mentors with different background experiences and from different STEM fields. As this work is sponsored by the Navy, these first four mentors will all be related to Navy STEM careers. In this paper, we primarily discuss the work already completed, including the process of capturing a mentor, technical design, and lessons learned so far. However, some discussion will also present issues related to scaling up this approach, both to a relatively small set of mentors and to wider use.

BACKGROUND: THEORETICAL UNDERPINNINGS

Careers pathways are strongly influenced by an individual's background and goals (both career and personal). Judge and Watanabe (1994) found that individuals' careers conflict with family relationships and personal goals for 68% of the population (known as spillover effects), which appears to have risen since that time (Nomaguchi, 2009).

Conflicts are caused by a complex interaction of work hours, salary, respect, stress, job satisfaction, and non-work satisfaction. Worse, people are bad at predicting the needs and feelings of their future selves (Wilson & Gilbert, 2005). Informational interviews allow students to model their goals on the trajectories of existing professionals. While personal histories are not perfect (e.g., prone to selection bias), they provide an authentic perspective on a career. Particularly for Navy-relevant STEM careers, opportunities for enlisting or advancing in less-common fields can be opaque to the typical student (or even many sailors within the Navy).

Informational interviews, mentoring consortia, and related ways of providing information on job realities have been effective for increasing motivation, engagement, and career self-efficacy (Herman, 2010; Krieshok et al., 1998; Murrow-Taylor et al., 1999). Unfortunately, these programs scale poorly: it is impractical for every student to have access to diverse mentors from a variety of careers. Worse, mentorship programs may reproduce entrenched STEM demographics, which are disproportionately white and (except for life sciences and psychology) male (Lehming et al., 2013). This makes it difficult for students to find a role model who has similar life experiences (e.g., ethnicity, socio-economic status, gender), despite the fact that these factors impact students' career identity and pathways.

Virtual agents have been particularly effective at communicating individuals' experiences to a wide range of people, by providing interactive dialogue and narrative (Bickmore et al., 2010; Swartout et al., 2006; Swartout et al., 2013; Traum et al., 2015). In some contexts, people self-disclose more information to agents, since they do not feel judged (Gratch et al., 2014). By modeling individual Navy STEM professionals, our agents should facilitate role modeling (Bandura, 1986) and anchoring goals to career trajectories of individuals already in that job (Kuder, 1980).

In terms of virtual agents, this project directly builds on findings and technologies from two prior projects: New Dimensions in Testimony (NDT; Fig 1) and Personal Assistant for Life-Long Learning (PAL3; Figure 2) project. Both projects implement unique capabilities for designing virtual agents that support learning. The NDT project serves as a model for how to build a virtual agent who communicates their life experiences (or, in this case, career experiences). NDT built a novel approach called "time-offset interaction" to emulate a realistic conversation with a specific individual, rather than a generic archetype (Artstein et al., 2014; Traum et al., 2015). NDT is intended to maintain and share the experiences of Holocaust survivors (Figure 1). In NDT, video recordings were made of a Holocaust survivor answering questions that a typical museum goer might ask. Subsequently, NDT would present the recorded survivor on a large screen. When a visitor asked a question, NDT would use speech recognition and natural language processing algorithms to determine what the question was most likely about and play the pre-recorded answer. Artstein et al. (2014) found that approximately 1400 answers could accurately answer 70% of questions in an open-ended conversation, and iterative improvement allowed under 2000 answers to cover 95% of all questions in a highly-natural conversation. Conversations with NDT are immersive and moving: visitors to the exhibit have been moved to tears or even to apologize to the virtual version of Pinchas Gutter, the first Holocaust survivor who has been modeled.

Figure 1. NDT at the Illinois Holocaust Museum

Figure 2. PAL3 Home Screen

While this level of breadth is cost-prohibitive for this project due to the goal to scale up to many agents, effective conversation should be possible with a smaller answer set by focusing more tightly on career-relevant dialogue. One goal of this project is to identify optimizations that maintain effective interactions, while lowering costs closer to a point where it might be possible to model hundreds of mentors, rather than just a handful, while still retaining the immersion and feeling of interacting with a real-life person. By modeling individual STEM professionals, our agents should facilitate role modeling and anchoring goals to career trajectories of individuals already in that job (Bandura, 1986). This approach can also help reach under-represented populations in STEM (e.g., minorities and women), by modeling a diverse set of role models as virtual agents.

The PAL3 project provides a platform (PAL3) and an interactive guide for learning (Pal; the robot in Fig. 2). The long-term goal of PAL3 is to track learning records for Navy sailors and use these resources to provide personalized learning recommendations based on their performance and career path (e.g., future postings). PAL3 will be used to guide students to mentors and career resources related to those mentors, while indirectly assessing students' career interests (e.g., such as Holland codes that assess dimensions of career interests; Holland, 1973) and using these to customize recommendations. To keep learners returning to the system, PAL3 has a variety of factors to increase engagement (Swartout et al., 2016). These include witty dialogue for the Pal guide, open learner models, and unlocking customizations for Pal. In a 2015 demo at I/ITSEC, K-12 students found the system highly interesting and engaging, which was part of the inspiration to adapt this platform to deliver STEM career interventions. MentorPal is being embedded into PAL3 as a new resource type that can provide an interactive virtual mentor. Since these MentorPal agents can be recommended as part of a topic, they can also be complemented with supporting resources to assess career interests and to help students understand career activities and expectations (e.g., example tasks).

TECHNICAL APPROACH

This paper will discuss the process of creating the first such virtual STEM mentor prototype, including the development of an extensive mentoring question bank (approximately 500 questions); key mentoring topics that intersect STEM, DoD, and civilian life; techniques for cost-effective recording of remote mentors; and the process of training and verifying a natural language dialogue model for answering and suggesting career questions.

Mentoring Topics and Questions

The process of building a question bank had four core steps: accumulating questions, developing criteria for assessing question quality, ranking questions, and categorizing questions into topics. The criteria for selecting and ranking questions were based on our target audience: students of approximately high school age. This period of life represents the most common age when students set goals to either continue their education (e.g., college, vocational institutions), join the military, or enter the civilian workforce. Since such students cannot always independently generate effective questions, the question bank was intended to cover both common student questions and prompts students to ask deeper or more complex questions they may have not thought to ask.

To generate a candidate question bank, a set of five researchers on the project each generated approximately 50 unique questions (for an initial bank of 250). Adding to this, approximately 175 questions from publicly-available question banks on Navy careers and STEM careers were retrieved and added to the set of candidates (sources included: National Education Empowerment Foundation, Inc., 2006; Los Angeles Mission College, 2014; Hansen, 2016). A review of the questions removed duplicate questions and also added follow-up questions related to existing unique questions, which increased the total number of questions to over five hundred questions.

Criteria were developed to evaluate the questions, which considered three orthogonal facets: priority, expected frequency, and engagement. *Priority* represented the team's belief that the question was important to career decision-making (i.e., "should ask") rated on a scale from 0 (not important) to 2 (very important). *Expected frequency* represented the team's expectation about the odds of at least one student in a class asking a particular question in a 30-minute session, rated as *high*, *low*, or *never*. *Engagement* represented the likelihood that a question would generate an interesting response that was either emotional and/or narrative. Questions flagged as having both an emotional aspect and a narrative prompt (e.g., "How did you overcome your biggest failure in your career?") were considered the most engaging, while those with only one such flag were next-highest. These facets for engagement were based on findings from the NDT project using similar technology, which found that storytelling, emotional cues, and genuine commentary establishes engaging and memorable interactions (Traum et al., 2015).

These dimensions were used to rank the questions, after which we selected the first two-hundred questions to record over the course of five three-hour interview sessions. The selection rules for questions targeted questions that scored highest on any of the three different criteria (i.e., the maximum of the three dimensions, rather than the average). From a design standpoint, this is an important choice: the system needs to be able to answer common questions, even if they might not be highly informative or engaging (e.g., "What is STEM?"). Likewise, questions with low frequency and high priority represent those questions we deemed as having high information value: these questions would rarely be asked, but should be, as they could offer the most beneficial information to the target audience (e.g., "Who gave you the best advice about your career?"). Finally, prior experience has shown that question prompts that produce high engagement are worthwhile to build rapport even if rarely asked (e.g., "Tell me a funny story.").

While choosing the final two-hundred questions, candidate questions were categorized into a set of topics to assist identifying similar questions and to record related questions in the same session. The set of question topics is shown in Table 1. The question set was intended to cover a meaningful range of topics that enable responses to realistic input from the user. Questions that rated highly on all categories were recorded directly. For questions that were commonly asked, but otherwise low priority and low engagement, we added instructions to guide the mentor to provide useful knowledge on top of the basic answer to the question (e.g., the answer to "What is STEM?" includes what skills are used in the mentor's STEM field).

Table 1. Question topic categories, with an example question from each topic.

Category	Question
Advice	*Should I go to college before or after serving?*
Background	*Where did you grow up?*
Benefits	*How do the educational benefits of the military work?*
Challenges	*What is your strategy for overcoming hardships?*
Colleagues	*How do you and your coworkers socialize outside of work?*
Combat	*What if I want to join the military but I'm afraid of going into battle?*
Conflict	*Tell me about a time when your priorities conflicted with that of management.*
Culture	*What do you find unique about your career field?*
Decisions	*What branch of the military should I go into?*
Demographics	*What is the gender mix of your workplace?*
Education	*Should I go to college before or after serving?*
Ethics	*Is sexual harassment common in the military?*
Failure	*What are the failures for which you are the most unhappy?*
Fit	*What happens if I don't enjoy this kind of work?*
Future	*What sorts of changes are occurring in your occupation?*
Growth/learning	*What range of tasks and opportunities does this field allow me to grow?*
Impact	*How can your work change the world?*
Jargon	*What does EMC stand for?*
JobSpecific	*Though you were an Electrician's Mate, did you work on things other than electronics?*
Lifestyle	*What do you do during your off time when you aren't deployed?*
Management	*How satisfied have you been with your bosses in your career?*
Mentor	*Who are the best people to talk to when making a career choice?*
Military	*What was the coolest thing you did in the Navy?*
Miscellaneous	*What's up?*
Misconceptions	*What are some of the common misconceptions about your field?*
Money	*How much money do you earn working for the Navy?*
Motivation	*Why do you love your job?*
Opportunities	*Where will a STEM degree take me?*
Project	*How can you help me?*
Priorities	*What priorities would you recommend for someone entering this field?*
Recognition	*Can you explain the ribbon system?*
Resources	*Where should I go to find out more information?*
Risk	*What were some of your fears entering college and your career?*
Skills	*Do you have to be good at math or science to go into STEM?*
STEM	*What is something most people don't know about computer science?*
Training	*What is the training progression of the Navy and what is it like?*
Travel	*How much travel did you do during active duty?*

Interview Preparation and Process

Since the mentor agent consists of video clips of a real-life mentor, a series of recordings needed to be conducted. Our equipment choice was driven by two goals: 1) to create an affordable, repeatable process that does not significantly impair the quality of recordings, and 2) to enable a remote mentor to record their responses (since mentors might be stationed in many locations around the world). The total equipment cost was $235, excluding the standard desktop computer used to record sessions, which was a standard computer lab machine. The specific items were a Microsoft LifeCam webcam (~$55), a Blue SnowBall microphone (~$80), two umbrella photography lights, a backdrop stand with sheets (black, white, and green; ~$120), and free software programs for video chat (Skype) and video recording (Open Broadcast Software (OBS), open source on obsproject.com). In theory, the lighting and backdrops would be optional for a location with excellent and consistent lighting and backdrops. However, the lighting setup is still recommended, since good lighting is seldom the case in practice and changes in lighting can cause uneven images when video clips are sequenced. This specific equipment and software is not unique—nearly any webcams, microphones, and recording software should be effective. However, these are noted as a reference point for the level of recording quality and cost used for this project.

Prior to sending this equipment to the first mentor, we tested the equipment in a similar environment, and developed set up instructions suitable for a variety of users. The setup required setting up the webcam, microphone, and backdrop, installing OBS Studio and the software for the webcam, and changing a few settings in OBS Studio for the desired output format and video/audio sources. For this output, we found that it was not necessary or even ideal to save the highest quality footage, as the lower quality format saved storage space and provided all the resolution and capacity necessary for effective replay within the final tablet application. These steps did not pose any significant difficulties in the setup process, and though the instructions were a good resource, our first mentor was able to figure out most of the setup without the assistance.

The first mentor selected for this project was chosen for his Navy STEM background as a nuclear Electrician's Mate. An addition to his STEM expertise, he is an engaging story-teller, he had prior experience as a student researcher at ICT when he pursued his computer science degree, and he represents a second STEM field in the civilian sector (Computer Science). This experience developing computer systems helped him suggest improvements to the recording and interview process. As such, he represented a good first mentor who could both record dialogues but also collaboratively improve the recording process itself. For example, he provided cues to say more on a topic and suggested approaches to upload videos to the research team. However, this also means that our first mentor represents a best-case scenario from a technical standpoint, as his STEM expertise includes strong computer skills. As such, minor revisions to instructions will likely occur with future mentors. As our first mentor, he will be referred to as "Mentor 1" through the remainder of the document.

Mentor 1 was in Japan during the recording period, so the recording process itself proceeded via a common video chat tool for distant communication (Skype). Since the recording was done by software and equipment on the mentor's computer, quality of the Skype footage was irrelevant given that it provided uninterrupted communication for the interview, which it did quite effectively. The preparation process immediately preceding the actual sessions included turning on the two studio light umbrellas, turning on the mentor's personal computer and ensuring that the video and audio from the mic and webcam were working correctly, and finally opening OBS Studio and hitting record before starting the interview. During breaks in the interview session, the mentor would occasionally be asked to make sure that the recording was present on their machine and could be played. During each recording session, Mentor 1 needed to wear the exact same clothing and have the same background. By recording Mentor 1 in his uniform, consistency across recording sessions was very simple compared to civilian recording subjects, who can problematically change haircuts or wardrobe across multiple months.

Before recording each session, we found it pertinent to prepare the speaker to respond to the questions that we would be asking in a way that would be effective when speaking to our target audience: pre-college students with little knowledge of STEM or the military. To cover the contextual portion of successful storytelling, we prepared a small instruction to recite before beginning each session:

> *"Imagine you are standing at the front of a room talking to a crowd of high school students. The students know very little about careers, and even less about STEM or the military. They will be asking you questions to help them prepare for their futures, and to decide which career path they would like to take, whether it be STEM, military, or both. Try to answer the questions with this context in mind, and feel free to pause before answering or give multiple answers. We may repeat questions asked from previous sessions."*

This introduction was effective in setting the tone for the interview, but we found that as the three-hour session went on, the mentor would often start to assume prior knowledge. He would answer the questions in a way that assumed the audience had previous background knowledge regarding the matters about which he was speaking. To solve this problem, when returning from breaks throughout the progression of the session, we would reiterate the necessity of speaking to an audience of pre-college students with little knowledge of STEM or military careers.

In addition to the scheduled questions for a recording session, follow-up questions were identified by the experimenters during the session and time was allocated to record answers to those questions. This was necessary because when the final system interacts with a human, that user will ask follow-up questions based on the content of the mentor's answers. This is the essential quality of effective virtual interaction that balances the story designed by an author with the emergent story during a given user's conversation path (Lee, Oh, & Woo, 2005). Therefore, it was important to identify follow-up questions and predict various paths taken by users preceding and following particular mentor responses. A set of follow-ups were predicted beforehand and asked directly after their source questions. However, unpredictable follow-ups also surface based on a mentor's unique and spontaneous response. To handle this, most of the interviews were held with several people in attendance, allowing for real-time follow-up questions and generating a broader question base from the group. Though this was effective for active generation of further inquiries, having additional people present during interviews also caused lengthier discussion between questions, producing fewer questions per session (an average of 50 non-follow-up questions per 3-hour session), and sometimes breaking the flow for the mentor (e.g., hesitance or uncertainty from multiple follow-ups that reduced the conversational tone of the interview). For Mentor 1, ease of speech tended to be higher with fewer people watching and increased with each successive recording session as the mentor became more comfortable. The pros and cons for different interview group sizes will be considered as further mentors are recorded.

To provide for delayed generation of follow ups or recognition of inadequate responses, it was essential to rerecord questions with supplementary or edited responses. Despite our desire for authentic statements in a conversation, it was sometimes necessary to request a second response due to overly long answers (beyond an established cutoff of 3 minutes), answers that refer to prior answers in a session (e.g., unresolved pronouns), stuttering or poorly worded responses, inappropriate answers, or answers that are time-dependent and may be false in the future upon use of the system. Particularly, for some of the commonly-asked questions, it was sometimes necessary to outline information for responses, since some questions were about resources or processes about which the mentor did not have thorough knowledge. In the case of delayed follow-ups or questions repeats, a debrief was first given to provide context to the follow-up or the desired details and reason for a repeated question. By providing this background, we productively improved the responses despite the questions coming at disparate times.

Dialogue Model Development

In the first stage of developing the dialogue model, Mentor 1 was interviewed, and 298 unique questions were collected over five interviews (this number includes follow-ups). For each question, 3-10 paraphrases were generated, as a question can be asked in many ways and the system should be able to handle the variations. For example, "What do you do in the Navy?" can be asked as "What are your responsibilities in the Navy?", "What do you do in your job?" and similar questions. The goal is to provide as many variations as possible, so that the machine learning algorithms can identify good matches to a larger set of possible questions. During the original sessions, each interview session was recorded as a single large video. In the post processing phase, each interview video was split into smaller files based on annotations for the start and stop of the mentor response, so that each file contains a single answer. A commercial cloud-based engine was used to obtain initial transcriptions of the answer (IBM Watson Speech-to-Text (STT); www.ibm.com/watson/developercloud/text-to-speech.html). This specific engine was chosen because it allows opting out of commercial data analysis, which can be relevant for use in studies where student input is processed. Despite the STT, there is a fair amount of manual work involved, as the STT engine does not provide perfect transcripts. It fails to understand domain-specific jargon (Navy terms), it works poorly when speech has an accent, and it fails when speaking quickly (very brief pauses between words). Once accurate transcriptions for each answer are obtained, a corpus was generated with the questions, paraphrases, answers, and their topics from the set described in a previous section. These topics are helpful in finding out the answer for a question asked by the user. The flow of the data preparation process is shown in Figure 3.

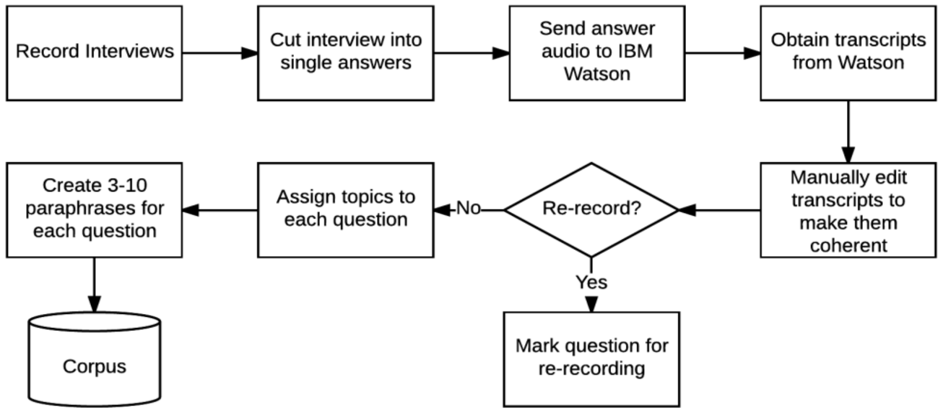

Figure 3. Process for Creating Corpus of Question-Answer Data for a Mentor

This corpus was processed to generate a dialogue model with three components. The first is NPCEditor (Leuski and Traum, 2011), a program developed at USC ICT for building a spoken language interface in virtual humans that engage with users. The second component is a domain-specific neural network classifier, which gives a suitable answer to a question asked by the user. The third component is a script that combines the first and second components by choosing the best answer to provide to the user. The script also interacts with the GUI and handles special events such as the user not asking a question for a while, the user asking an irrelevant question, the user repeating questions, etc. These events have specific responses that are triggered. The script is the critical component of the system as it interfaces with the GUI. For a given question from the user, NPCEditor and the classifier will each provide an answer which they think is the best response and the confidence score for that answer. If NPCEditor is confident of the answer it provides, indicated by a high score, then that answer is chosen. Otherwise, the answer provided by the classifier is used. There are also some situations when NPCEditor fails to provide an answer. In this scenario, the answer from the classifier is used. Similarly, NPCEditor is also used to handle certain special signals (e.g., identifying repeated questions) that the domain-specific classifier does not address. The effectiveness of this model is currently being evaluated through initial user testing, with initial indications finding that while it has reasonable topic coverage, more questions will need to be recorded to reach a suitable level of coverage for common questions (currently approximately 50% of new user questions do not map to a known answer). While this is not surprising since the NDT project required 1400 questions to reach 70% coverage, it does indicate that the model can only work as well as the responses available.

User Interaction and Interface

The goal of the user interface for MentorPal is to put focus on the mentor, rather than the interface. This called for a minimal interface with obvious affordances for how to interact with the system (Mueller, D., & Strohmeier, S., 2011), emulating the type of low-fidelity chat you would see in a video conferencing solution (e.g., Skype, Google Hangouts, etc.). This is particularly relevant to a younger audience, since a simple user interface eliminates distractions (i.e., clicking buttons just to see what they do). As shown in Figure 4, the user interface is intended to be run inside a tablet application, focuses on a dominant centered screen within which the mentor responses are played back in response to user input. Below the video screen area is a text input field that records questions and sends the transcribed text to the classifier. As an alternate input, a voice-input button is also present (the button on the far left of the transcript area). Directly above the text input and below the video display, the field of transcriptions is shown for the questions and responses during that conversation session. On the top of the interface, above the video screen, is basic information about the mentor. As this system is built in to PAL3, it is a Unity-based application.

When a question is submitted by the user, a call to the question classifier is made which then selects the best match and returns the appropriate video clip, which is then loaded and played on the application with minimal delay (i.e., like a real-time conversation), while also loading the transcribed text for the clip into the scroll box. The integration of voice and text emulate a chat structure, establishing a familiar setting for interaction with mentors. This chat structure also reduces expectations for the video fidelity for the mentor (i.e., the feeling of a remote video chat, where artifacts and occasional jitters are accepted). Despite the simple interface, a number of substantial design decisions were made that might influence future systems following this model.

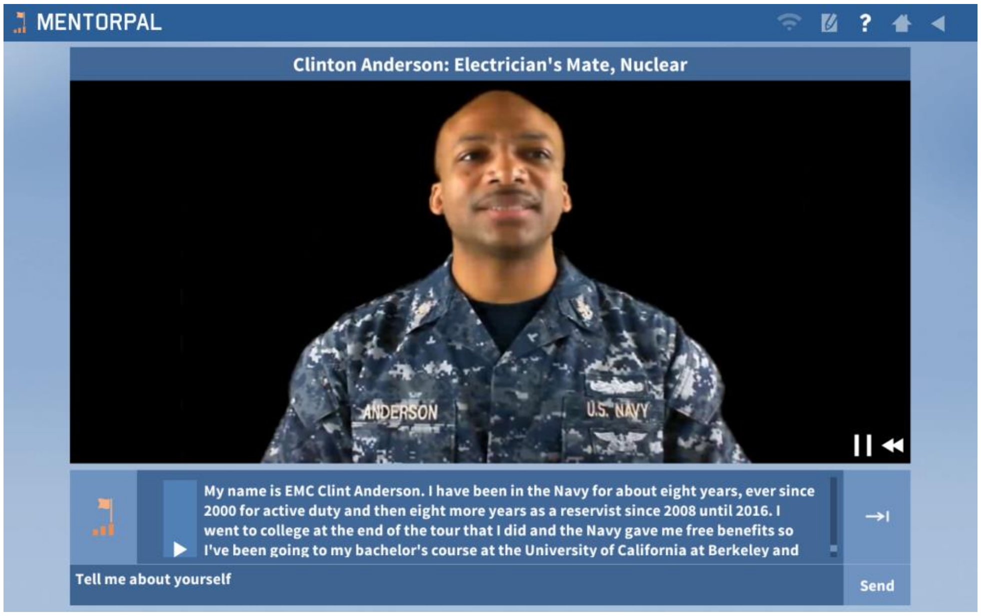

Figure 4. GUI prototype with video playback, transcription scroll box, and text input.

First, the decision to display the user's speech-to-text results on a transcript is non-trivial. In general, this is because natural-language dialogue systems can often respond well despite poor transcriptions of what a user said to a system (D'Mello, Dowell, & Graesser, 2011). So then, if the user's text-to-speech results are not shown, in some cases the system might technically receive a poorly-transcribed question but still respond with an answer that seems reasonable. However, as soon as the text-to-speech input is shown, the user can note that their question was not truly understood (i.e., two points of failure). One section of the team favored a reliable display of the user's received input, to help users repeat their question if it was not understood properly. If such problems happened frequently, students would have no real recourse if the voice-recognition software did not understand them. The alternate camp favored a dialogue-based experience for the user, to avoid distracting them. This view also raised the point that text presentation might make users more fixated on monitoring the system's accuracy and treating the dialogue as a machine, even to the point of trying to challenge it with unfamiliar or alternatively pronounced terms. Since the implementation itself is trivial, this issue was treated as a research question and the display has been tentatively included. As usability data is collected, early users will be asked about the impact of this feature.

A similar issue was raised in the choice of including an ability to replay, pause, or rewind the mentor's answer. Those in favor saw it as a good review process, should the user want to hear the answer again (e.g. they were interrupted, missed the point, or wanted to share it with a nearby student). Those favoring an experience based on "conversational flow" objected to this as an unnatural action. In conversation, one would normally just ask the person to repeat himself or rephrase what he had said. The opposing view also illuminated that this would provide a substantial departure from a normal video chat. Further, there was the issue of how this feature might be implemented on platforms with different constraints (e.g., smartphones). As with the display of speech-to-text on the transcript, these features were included but may be removed based on user experience data.

Another implementation issue faced by the team was how the mentor was displayed in terms of size, shape, and prominence. The use of live clips for responses requires a precise and disciplined return to a common "rest position" at the beginning and at the end of each clip to avoid discontinuities between clips. This situation would clearly be exacerbated if the mentor's screen showed their entire body (e.g., leg positions tend to move significantly). The team further discussed the desirability of showing enough of the clothing of the individual to convey their rank and military service. Those in favor of a larger coverage (waist-up) argued that body language and gesticulation are important communicative resources that would be curtailed by a harshly truncated image. The compromise was made to use a classic "head and shoulders" image frame, sufficient to show the mentor's military decorations and campaign ribbons (Figures 5a and 5b).

Figure 5a. Head and shoulders video display with patches visible.

Figure 5b. Head and shoulders with gestures.

With the video playback comprising the most significant portion of the screen, we chose to record mentors from chest up in order to provide a personal experience, as if the user was truly sitting with the mentor inquiring about his or her future. Although a closer view provides more emotion due to the detail of facial expressions and head movement, lower body movement such as hand gestures are often out of view. To make up for this, we emphasized to Mentor 1 the importance of animation, visible hand gestures, and emotional oration (Figure 5b). The mentor positioned the monitor in a way that would allow him to maintain eye contact with the user as much as possible while still gesturing naturally. They were then instructed to find a comfortable position which they would return to at the beginning and end of each response. While this process took some tuning on the initial recording session, returning to the rest position quickly became natural for Mentor 1, leading to relatively uniform recordings even across multiple sessions. Future work will identify if later mentors find these principles equally easy to follow.

DISCUSSION: IMPLICATIONS FOR SCALING UP

There are a number of scale-up considerations for the MentorPal program, both in the small scale (the next 3 mentors to be recorded during the next year) and for potential large-scale use in the long term. These scale-up considerations fall into three main categories: cost, expertise, and mentors. As noted earlier, cost barriers appear to be relatively low. The equipment and software required to record mentors is less than $250 and could be re-used for multiple mentors at the same site. Likewise, as a tablet-based system, the main technical barrier to scale up would be video hosting, which some providers are selling unlimited bandwidth at $50/month for 5 TB of videos (which would be enough for about 250 mentors at the recorded video bitrate which will ultimately be about 20 GB/mentor, and up to 500 mentors at a lower bitrate). So then, the hardware and cost barriers are fairly low for scaling up.

The expertise barriers to recording mentors have not yet been fully explored. In this paper, we note several subtle design choices and issues that should influence how future groups consider recording interactive mentors. These include setup requirements, question choice, priming statements to ensure that responses target the right audience, user interface design, and processes for follow-up questions and re-records. However, as only a single mentor has been recorded, the instructions and guidelines are not yet firm enough for any group with no expertise to easily replicate the process of recording a mentor and training a dialogue model. As more mentors are recorded, it is a goal of this project to refine this process and identify the documentation and automation that would be required to build a "MentorPal Recorder" that would enable a group with no AI or advanced technical expertise to record a mentor and generate an interactive agent. While this is not yet a fully-explored issue, we feel that it is entirely possible to make rapid recording and playing-back such mentors something that could reach the level of a Government-off-the-shelf (GOTS) technology. A significant unanswered question is how much of the dialogue model and question bank can be transferred to a new dialogue model (e.g., can a variant of the same model be used?). A key issue for reducing the expertise barriers is that authors cannot be required to manually update the dialogue model. Another question is how to update mentor responses if knowledge changes. Currently, this is handled by avoiding particularly topical responses, but some mentors might require archiving outdated responses or updating recordings in the event of large changes in a career path.

More generally, for the MentorPal approach to be effective, the mentors must be both *useful* and *engaging*. While the pathway toward a scalable technology appears to be mostly solvable, the more fundamental issues of selecting mentors to record, evaluating mentor quality, and helping match learners to the right mentor are substantially more complex. Selecting mentors has two distinct levels to consider: mentor quality and mentor salience. From one point of view, there are particular qualities that determine a good mentor (e.g., useful information, engagement, desire and interest to serve as a mentor). This perspective represents a mentor as an authority in a career, such as a teacher. However, the alternate perspective considers a mentor as a counterpart who shares and compares a series of life experiences and career trajectory with a student. In this second case, an ideal mentor would be one whose experiences are most comparable the student, so that they can share the specific opportunities and pitfalls that exist (i.e., similar trajectories). Both perspectives have value, which implies that from the standpoint of a group of mentors, the ideal would be that all mentors are individually good mentors but that they also cover a wide range of experiences so that different students can understand how to relate career advice to their own life.

In terms of traits for an effective mentor to deliver as an agent, three distinct factors were identified: ability (career knowledge, career growth, perception, self-reflection), interest (mentor at heart, passionate about career, committed), and engagement (good to talk to, charismatic). Rose (2005) and Heathfield (2017) inventoried some of these universal traits, but specializing on Navy and STEM careers for students required specializing many of these traits. For this specific project, mentors require sufficient experience in the germane professional field to warrant credibility, a familiarity with the military service, an openness to discuss both failures and successes, and an authentic voice that places how their STEM and Navy careers relate to their larger life experiences. Engagement is the subjective ability to visually and verbally convey all the attributes in the first set to young adults, such as screen presence, conversational style, and relevance of the mentor's life experiences to the students. As noted, narrative/storytelling and ability to evoke emotions was valued in mentors, since these are known to improve both time-on-task and retention of information (Graesser, Halpern, Hakel, 2008). Interest was easiest to articulate: the mentor's interest in the project, combined with their prior desire and experience to act as mentors and leaders. Mentors for this project must devote on the order of twenty or more hours of questioning in a disciplined environment, where they must carefully return to a "resting position," and respond frankly to deep questions about their career and their life. This dedication is particularly important because if a mentor withdraws before sufficient footage is collected, their question coverage will likely be too poor to use for an interactive agent.

Table 2 Mentor Career Expertise and Proxies for Life Experiences

Name	STEM Expertise	Career / Rating	Navy Status	Career State	Gndr	Age	Ethnicity	SES Back-ground	Locale Back-ground	ICT POC Person	Notes	Con-tacted?	Selected ?
A. Aaaa	Electrician / CS	Electricians Mate - Nuclear	Reserve (Chief Petty Officer of a 35-member Reserve Unit,)	Mid	M	30	African-American	Mid?	Varied: Alabama, Michigan, Los Angeles	BN	Former student worker at ICT, involved in PAL3 project. Left Navy to pursue game design at USC, started college via GI Bill. Still active as a Navy reserve and passionate about better technology in the Navy (won a Navy essay award about it).	Y	Y
B. Bbbb	SWO	SWO	Retired CAPT	Retired	M	50	European	Upper	New York, California,	DMD	Was Dean of Engineering at NPS	N	
C. Cccc	Intelligence	CTICS	Retired CPO	Retired	M	65	European	Middle Class	Colorado, international	DMD	Very articulate, can speak to many technical areas that are part of cryptology in Navy. Lots of overseas experience. BA from Cal State Long Beach	N	
D.Dddd	Electronics and CS	LCDR, Submarine Service	Retired LCDR	Retired	M	65	European	Upper	Colorado, North Carolina	DMD	Knows EE, CS, Nuc Power, has taught, has put two kids through college	Y	
E.Eeee	Civil Engineering/ Cryptologist	Sgt USMC/ Construction/ Safety Officer	Honorable Discharge/USMC	CADRE	M	68	European	Middle Class	New Mexico, Minnesota	DMD	Construction Safety really technical, must know science and math, read and interpret technical specifications	Y	

Of the traits of a mentor, relevance is the most subjective: it is difficult to know whose life experiences will be most valuable to a younger person considering the same career. If we think of life as a series of decision points, the most relevant mentor would likely be an individual who encountered similar choices, with similar values, opportunities, and risks. Rough proxies for such similarity include basic demographics such as gender (Gray & Goregaokar, 2010), ethnicity (Cabral & Smith, 2011), life history (*e.g.* rural or urban; geographic location), age (Finkelstein et al., 2003), and socio-economic status. To evaluate options for potential STEM mentors for the next MentorPal phase, a table of potential mentors is being developed to help balance both the career expertise and the relevance factors so that different careers and backgrounds can be covered. A subset of this table (anonymized) is shown in Table 2.

CONCLUSIONS

In summary, the goal is not to find someone who speaks for every STEM or Navy professional, but to find someone who shares a relatable experience that will help a novice understand if they are a good fit or a poor fit for a certain STEM field. This final point is key: the goal of the MentorPal mentors is not to persuade students to do a certain career, but instead to help them increase the realism of their career goals: helping them visualize what life is like in a certain career and what kind of steps are needed to get there (Brown et al., 2003). With only four mentors for the next prototype, this means that only one path can be shown for each STEM career represented. In a larger scale mentoring situation, a student or sailor would be able to talk to multiple mentors who experienced the same career path. For example, the US Naval Academy Alumni Mentoring Program (www.usna.com/AMP) tries to provide mentees four mentors for key decision points (e.g., leaving the Navy): two that made each decision, where one was happy with their choice and the other was less satisfied. A large-scale version of a program like MentorPal would ideally be able to provide those types of alternate viewpoints and experiences. It could also approach many branches of both military service and civilian careers, both in STEM fields and in other growing careers.

Compared to such real-life networks, interactive agents have advantages that make them complementary to traditional mentoring techniques. The unique advantage of MentorPal is amplification. Particularly for new careers or persons of new backgrounds in a career (e.g., enlisted females on submarines), it would be effectively impossible for a small set of trailblazers to have face-to-face talks with much larger future cohorts. More generally, a system like MentorPal should scale up well, since running mentor video clips for a year would be on-par with a single in-person trip by a mentor to meet students outside of their region. The clear drawback is that while MentorPal will offer a personalized conversation with a student, it will not offer a truly personal conversation or ongoing mentoring relationship as would be experienced with a real-life mentor, so MentorPal offers greater breadth, but less depth.

Despite these trade-offs, this work has so far found that the costs for scaling up such mentors might be lower than expected in terms of hardware and software costs. However, further work is still necessary to demonstrate that dialogue models can be easily retrained on a new mentor without requiring any specialized skills. Finally, further fundamental research is required to systematically establish criteria for selecting a set of mentors who can speak to the experiences of learners interested in different STEM fields and different domains. As usability studies and efficacy data is collected, this project will look to identify factors that determine effective mentor-mentee pairs. Such match-making insight would be a benefit both for computer-based and traditional mentoring.

ACKNOWLEDGEMENTS

This work was sponsored by Navy grant N00014-16-R-FO03, and uses technology from multiple grants that have supported NPCEditor and PAL3, including Army contract W911NF-14-D-0005. The views stated in this paper are the responsibility of the authors alone and do not reflect the opinions or positions of sponsoring organizations.

REFERENCES

Artstein, R., Traum, D., Alexander, O., Leuski, A., Jones, A., Georgila, K., ... & Smith, S. (2014). Time-offset interaction with a Holocaust survivor. In *ACM Conference on Intelligent User Interfaces (IUI) 2014* (pp. 163-168). ACM Press.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall

Bickmore, T., Schulman, D., & Yin, L. (2010). Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence*, *24*(6), 648-666.

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.

Brown, S. D., Ryan Krane, N. E., Brecheisen, J., Castelino, P., Budisin, I., Miller, M. & Edens, L. (2003). Critical ingredients of career choice interventions: More analyses and new hypotheses. *Journal of Vocational Behavior, 62*, 411-428.

Cabral, R., & Smith, T . B. (2011). Racial/ethnic matching of clients and therapists in mental health services: A meta-analytic review of preferences, perceptions, and outcomes. *Journal of Counseling Psychology.*

D'Mello, S. K., Dowell, N., & Graesser, A. C. (2011). Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology: Applied*, *17*(1), 1-17.

Finkelstein, L. M., Allen, T. D., & Rhoton, L. A. (2003). An examination of the role of age in mentoring relationships. *Group & Organization Management*, 28(2), 249-281.

Graesser, A.C., Halpern, D.F., Hakel, M. (2008). *25 Principles of Learning.* Washington, DC: Taskforce on Lifelong Learning at Work and at Home.

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. In *Autonomous agents and multi-agent systems (AAMAS) 2014 Conference* (pp. 85-92). International Foundation for Autonomous Agents and Multiagent Systems.

Gray, D. E., & Goregaokar, H. (2010). Choosing an executive coach: The influence of gender on the coach-matching process. *Management Learning*, 41(5), 525-544.

Hansen, R. S. (2016). *200 Informational Interview Questions*. Retrieved Fall, 2016, from https://www.livecareer.com/quintessential/informational-interview-questions

Heathfield, S.M. (2017). 15 Characteristics of a Successful Mentor, The Balance. Retrieved from the internet on 01 May 2017 from: https://www.thebalance.com/top-characteristics-of-a-successful-mentor-1917831

Herman, S. (2010). Career HOPES: An Internet-delivered career development intervention. *Computers in Human Behavior*, *26*(3), 339-344.

Holland, John L. *Making vocational choices: a theory of careers.* Englewood Cliffs: Prentice-Hall, 1973.

Judge, T. A., & Watanabe, S. (1994). Individual differences in the nature of the relationship between job and life satisfaction. *Journal of Occupational and Organizational Psychology*, *67*(2), 101-107.

Nomaguchi, K. M. (2009). Change in Work-Family Conflict Among Employed Parents Between 1977 and 1997. *Journal of Marriage and Family*, *71*(1), 15-32.

Krieshok, T. S. (1998). An anti-introspectivist view of career decision making. *The Career Development Quarterly*, *46*(3), 210-229

Kuder, F. (1980). Person matching. *Educational and Psychological Measurement, 40*, 1-8.

Lee, Y., Oh, S., & Woo, W. (2005). A context-based storytelling with a responsive multimedia system (RMS). In International Conference on Virtual Storytelling (pp. 12-21). Springer Berlin Heidelberg.

Lehming, R., Gawalt, J., Cohen, S., & Bell, R. (2013). Women, Minorities, and Persons with Disabilities in Science and Engineering: 2013. *National Science Foundation, Arlington, VA, USA, Rep*, 13-304.

Leuski, A., & Traum, D. (2011). NPCEditor: Creating virtual human dialogue using information retrieval techniques. *Ai Magazine*, *32*(2), 42-56.

Los Angeles Mission College *(2014). STEM E-Mentoring Program Mentee Manual*. Retrieved Fall, 2016, from, www.lamission.edu/estudentservices/wp-content/uploads/2014/08/Manual-stem-mentee7-22-14.pdf

Mueller, D., & Strohmeier, S. (2011). Design characteristics of virtual learning environments: State of research. *Computers and Education*, *57*(4), 2505-2516. doi:10.1016/j.compedu.2011.06.017

Murrow-Taylor, C., Foltz, B. M., Ellis, M. R., & Culbertson, K. (1999). A multicultural career fair for elementary school students. *Professional School Counseling*, *2*(3), 241-244.

National Education Empowerment Foundation, Inc. (2006). *Some Questions You Can Ask People About Their Career*. (2006). Retrieved Fall 2016, from www.collegecareerlifeplanning.com/Documents/ 4%20Career%20Planning/ j%20Networking%20Internships/Questions%20to%20Ask.pdf

Rose, G. L. (2005). Group Differences in Graduate Students' Concepts of the Ideal Mentor. *Research in Higher Education*, *46*(1), 53-80.

Swartout, W. R., Gratch, J., Hill Jr, R. W., Hovy, E., Marsella, S., Rickel, J., & Traum, D. (2006). Toward virtual humans. *AI Magazine*, *27*(2), 96.

Swartout, B., Nye, B. D., Hartholt, A., Reilly, A., Graesser, A. C., VanLehn, K., … & Rosenberg, M. (in press). Designing a Personal Assistant for Life-Long Learning (PAL3). *Florida Artificial Intelligence Research Society (FLAIRS) 2016 Conference*. AAAI Press.

Swartout, W, Artstein, R., Forbell, E., Foutz, S., Lane, H., Lange, B., Morie, J., Noren, D., Rizzo, S., Traum, D. (2013). Virtual Humans for Learning *AI Magazine, 34*(4), 13-30.

Traum, D., Jones, A., Hays, K., Maio, H., Alexander, O., Artstein, R., . . . & Haase, K. (2015). New Dimensions in Testimony: Digitally Preserving a Holocaust Survivor's Interactive Storytelling *Interactive Storytelling* (pp. 269-281): Springer International Publishing.

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.

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: knowing what to want. *Current Directions in Psychological Science*, *14*(3), 131-134.