The Effects of Interaction and Visual Fidelity on Learning Outcomes for a Virtual Pediatric Patient System

Toni Bloodworth Pence and Lauren Cairco Dukes
School of Computing
Clemson University
Clemson, South Carolina 29634
Email: tonib.pence@gmail.com and lcairco@gmail.com

Larry F. Hodges
School of Computing
Clemson University
Clemson, South Carolina 29634
Email: lfh@clemson.edu

Nancy K Meehan and Arlene Johnson
School of Nursing
Clemson University
Clemson, South Carolina 29634
Email:nmeehan@clemson.edu and aejohns@clemson.edu

Abstract—One of the most common clinical education methods for teaching patient interaction skills to nursing students is role-playing established scenarios with their classmates. Unfortunately, this is far from simulating real world experiences that they will soon face, and does not provide the immediate, impartial feedback necessary for interviewing skills development. We developed a system for Scaffolded Interviews Developed by Nurses In Education (SIDNIE) that supports baccalaureate nursing education by providing multiple guided interview practice sessions with virtual characters. During the development and evaluation of SIDNIE we realized the importance of determining the visual and interaction fidelity requirements necessary for proper learning. In this paper we report on two fidelity studies conducted with nursing students. The goal of the visual fidelity study was to determine if our virtual characters containing life-like animations would have an effect on learning or if we would get the same effect using a stationary image of our virtual environment. The second study focused on the interaction fidelity of our system and the goal was to determine if the interaction modality had an effect on the learning outcome. In particular we evaluated the effect of voice input as compared to a standard mouse-click input for question selection.

I. INTRODUCTION

Nursing students have limited opportunities for interaction with real patients, especially with pediatric patients, and often do not receive immediate and impartial feedback on their performance during their patient interaction. In order to provide alternative educational opportunities, we created a system for Scaffolded Interviews Developed by Nurses In Education (SIDNIE), which allows nursing students to interact with virtual characters acting as patients to practice their interviewing skills, while receiving guidance and feedback from a virtual nurse educator.

Experiential learning through simulation may help students develop the skills necessary for clinical practice and help develop the self-efficacy and critical thinking skills they need to provide the safest care possible. Also, this system has the potential to provide consistent experiences for all students with feedback and repetitive practice.

Since children are a vulnerable population, nurses must develop appropriate pediatric skills necessary to provide safe care to infants, children, and teens. Because nursing students have limited access to pediatric patients and may not have the opportunity to practice with children, we focus on a mother and child scenario.

Our system is intended as a tool to aid nursing faculty in providing scored practice opportunities for student nurses. During the development and previous usability evaluations of SIDNIE, we realized the importance of determining the visual and interaction fidelity requirements necessary for proper learning. In this paper, we discuss two different fidelity experiments and how they affect learning outcomes. We were interested in determining if low/high visual fidelity had an effect on the learning outcomes and how participants reacted to the system in each condition. We were also interested in determining if a natural language interface would have an effect on the learning outcome and how participants reacted to using it.

II. RELATED WORK

A. Simulation Learning and Patient Safety

The Institute of Medicine has recommended the use of simulation training as a method to improve health care delivery [1], and patient safety has been identified as one of the six competencies necessary to improve health care education [9]. Simulation training is an effective strategy to help promote safe clinical practices [3] and impacts the development of self-efficacy and judgment skills for nurses that are essential to provide the safest and most effective care possible [4]. Simulation learning that mimics real world scenarios is beneficial to nursing students and will provide standardized experiences in which students can practice problem solving techniques and clinical decision making abilities [4]. Five advantages to using simulation in nursing education have been identified: 1) providing opportunity for interactive learning without risk to patients, 2) boosting students’ self confidence and reducing anxiety in the practice setting, 3) allowing nursing students to practice clinical decision making, and critical thinking in a controlled environment, 4) allowing skills and procedures to be repeated until proficiency is reached, and 5) providing immediate feedback [5]. We designed SIDNIE specifically to
provide repetitive practice as well as the benefits of immediate feedback.

One of the current simulation exercises in the Clemson University School of Nursing involves students acting out written scenarios, where one student acts as the patient while the other acts as the nurse. SIDNIE can replace the student patient with a virtual patient, which is an artificially intelligent human representation that behaves similarly to a human patient under the same circumstances. Researchers have used virtual patients to teach communications skills to medical students, and students have rated the virtual patient experience as being as effective as a standardized patient (a paid actor) [4]. Medical students have also used virtual patients to help practice patient interviewing skills with a high level of immersion. Results indicate that using life-size virtual characters with speech recognition is useful in their education [6]. Adult virtual patients are fairly common, but virtual pediatric patients are rare. The use of virtual pediatric patients was first addressed in [5], where they were used for training and assessment for medical students. The scenario in this system included a mother and daughter, but the technology for creating virtual human behaviors had not focused on children’s behaviors, therefore the realism needed for this project was not available [5]. Even with the lack of realism, the results of their study were positive in that many of the participants stated that they gained valuable experience.

B. Communicating with Children and Patients

In addition to the challenges presented in displaying appropriate animations for children, the dynamics related to the nurse-family-child relationship are extensive due to the many factors that enter into this relationship, including ethnicity, age, culture, and illness. During an assessment, the nurse must obtain information (verbal and nonverbal) from the parent(s) and child, and observe any interactions between them [4]. Studies by pediatric experts have shown that the nurse-family-child relationship is heavily dependent upon effective communication, which is a skill that is developed through interaction with different kinds of pediatric patients and families [7]. Student nurses must be aware of the interactions that may negatively or positively affect their communication skills, therefore affecting the relationship. This study also showed that a positive nurse-family-child relationship will promote the health of the child, while a negative relationship may have a negative impact on the health of the child [7].

C. Interaction and Visual Fidelity

Effective communication is widely accepted by the nursing community as a key factor in patient satisfaction, recovery and compliance [8], which is why communication skills are of great importance in nursing education and practice. There are many methods to improve communication skills, one of which is the process of verbally speaking aloud. Many researchers in the fields of human-computer interaction, virtual environments and others have all studied the effects of natural interaction. A study done by Johnson et al. showed that virtual patient systems using natural interaction methods will facilitate effective teaching and training with medical students [9]. Stevens et al. reports on the development and initial testing of a virtual patient system that incorporates an interactive virtual clinic scenario and virtual instructor to teach history-taking skills [1]. Students are able to interact with the virtual patient and virtual instructor via speech recognition software, which gives the student a more natural method of interacting with the virtual patient. Many students in this study reported that they enjoyed talking to the virtual patients and that the level of realism was similar to working with a standardized patient [1].

Studies have shown that using virtual environments for simulation training can be a very effective method [10]. A study by [11] showed that a high fidelity virtual environment resulted in a faster traversal of a real world route versus training with a standard map representation. The study done by Johnson et al. also showed that virtual patient systems with highly detailed human models create a more realistic experience [9]. Gulz et al. finds that the most established effect of animated agents in educational systems is the potential they provide to make the experience more engaging for the student [12].

III. System Description

The SIDNIE system is designed to teach nursing students pediatric patient interview techniques by providing interview practice with guidance and feedback from a virtual agent named Sidnie. Sidnie is a male nurse that serves as our virtual nurse educator and he guides the user through several scaffolded practice opportunities and provides feedback on user choices. Our nursing collaborators developed a patient-nurse interaction scenario based on a five year old child with an earache. This scenario was used to generate the questions and answers present in the simulation. The user conducts an interview with the virtual patients by selecting questions from a preset list of questions developed by our nursing collaborators, and the virtual patients respond appropriately by performing speech and animations based on the question selected. The user can also view interview playbacks and score them as if they were an instructor, encouraging knowledge application to new situations.

One of SIDNIE’s novel aspects is that it provides criteria-based scoring on questions that students ask. Currently, SIDNIE scores student questions on two aspects: age-appropriateness, which means that if the question is addressed to the child, it uses words that the child will understand, or if the question requires more difficult phrases, it is directed towards the parent, and unbiasedness, which means that the question does not imply a certain answer or assume something to be true that the patient has not confirmed. Our nursing collaborators identified these attributes as important characteristics for questions in a successful pediatric patient interview. Each question choice presented to the student is scored within our database on the basis of these two criteria, allowing SIDNIE to give automated feedback and scoring on each question the student selects. Figure 1 shows what the user sees when interacting with the system.

The application runs on a standard desktop computer and a single monitor, and the user interacts with the system using the mouse. Our virtual environment was created using Blender [14] and then imported in Unity 3D [15] where most of the application is housed. We also used Blender [14] and Poser Pro [16] for the character development and animation. SAPI
SIDNIE allows the user to interview the virtual patients by selecting from a list of questions. The user can navigate the interface by selecting tabbed options. The virtual patients respond with speech and animations.

5.3 [17] was used for the text-to-speech and IVONA [19] was used for the two characters voices. Our application also has an underlying SQLite 3 [18] database that provides the flexibility to change scenarios, scoring criteria, and other data. For a more detailed description of SIDNIE including scenario development, system design, scaffolded level descriptions, and system implementation please refer to [13].

IV. EXPERIMENTS

A. Experiment Description

Here we report on two different fidelity experiments conducted on SIDNIE. We were interested in determining the visual and interaction fidelity requirements necessary for proper learning and measuring the learning outcomes in each experiment. Both of the experiments used the same version of the system and the same scenario: a mother with her five year old daughter, who has an earache.

The first study focused on the interaction fidelity of our system and the goal was to determine if the interaction modality had an effect on the learning outcome. In particular, we evaluated the effect of voice input as compared to a standard mouse-click input for question selection. In this experiment, both conditions viewed the characters with life-like animations and the conditions represented the different interaction modalities. The low-fidelity condition used mouse-click interaction to select questions to ask the virtual patients. For the high-fidelity condition, we used a Wizard-of-Oz approach to simulate voice recognition. The participant would read aloud the question they would like to select and the experimenter would select the question via key press, without the participant’s knowledge. An example of a student talking to the system is shown in Figure 2.

The goal of the visual fidelity study was to determine if virtual characters containing life-like animations would have an effect on learning as compared to a static image of the characters within the virtual environment (i.e., a screen-shot). There were two conditions, where participants in one condition interacted with animated virtual characters, while participants in the other condition only viewed a static image of the virtual characters within the virtual environment. They used the system via mouse-click to select the questions they wanted to ask the virtual patients. In this experiment, the low fidelity condition is the group containing the static image of the virtual environment, while the high fidelity condition is the group containing the life-like animations.

Our primary hypothesis in the interaction fidelity study was that participants would prefer the interaction method consisting of speech input and that the learning outcomes from that study would show higher scores for the condition including the speech input. Our primary hypothesis in the visual fidelity study was that the participants would prefer the system using the life-like characters with animations and that the learning
outcomes from that study would show higher scores for the high fidelity condition.

B. Experimental Procedures

In both experiments, the participants read and signed an informed consent and completed a demographics questionnaire. The demographic questionnaires included questions about previous computer usage, experience in health care, and exposure to virtual environments.

We then asked the participant to fill out an interview pre-questionnaire to act as a baseline so that we could measure the learning outcomes of using the SIDNIE system. The experimenter presented participants with a scenario of a mother who brings her four year old son into the doctor’s office for a stomachache. We listed the definitions for age appropriateness and unbiasedness that the SIDNIE system uses, then asked the participants to write up to five questions to ask the parent and child that were both age appropriate and unbiased. By asking the participants to write questions that met the criteria, we aimed to measure the participant’s learning on the middle level of learning in the revised Bloom’s taxonomy, “application” [21]. We also asked participants to score five questions on whether they were age appropriate and unbiased to gage learning on a higher level in the taxonomy, “evaluating”.

Next, we seated the participant at a desktop computer and began the SIDNIE system by taking the participant through tutorial screens, which instructed them on how to use the system. Once the participant was done with the tutorial screens, we asked the participant to conduct an interview with the patients to the best of their ability, and to let us know when they had finished. The SIDNIE system required that the nurse ask a correct (both age appropriate and unbiased) question from each of the 15 categories. SIDNIE gave feedback on each question selection, displaying on screen whether the question was age appropriate and unbiased. If the participant selected an incorrect question within a category, SIDNIE required that the participant kept trying until he or she selected a correct question before the participant could proceed to the next category. At the end of the scenario, participants received feedback on their overall performance in chart form, with a line for each question the participant asked during the interview that contained its scoring information for the criteria.

After the participant completed the interview, we asked them to fill out the interview post-questionnaire, which was identical to the interview pre-questionnaire so that we could measure learning outcomes. After completing the post-questionnaire, the participant filled out the System Usability Scale (SUS) [20] and was given the opportunity to provide any written feedback. The participant was also given a questionnaire on co-presence adapted from the Slater co-presence questionnaire found in [22]. Co-presence refers to the participants sense of being with another person, and may be interpreted as a measure of a character’s realism.

Finally, the participant completed a debriefing interview with the experimenter, where we asked questions about his or her opinion of the system and transcribed their answers.

C. Participants

Twenty-one students took part in the interaction fidelity study, with 11 participants in the high fidelity condition and 10 participants in the low fidelity condition. Their ages ranged from 22 to 50 years old (mean=27.62, sd=8.37). These students
TABLE I. STATISTICS FOR PRE- AND POST-TEST SCORES FOR QUESTIONS WRITTEN BY STUDENTS AND SCORED BY EXPERTS. IN THIS TABLE, SCORES ARE CALCULATED WITHIN EACH CONDITION AND WITHIN EACH EXPERIMENT. THE STATISTICAL TESTS MEASURE THE EFFECT OF THE CONDITION ON THE CHANGE IN TEST SCORES, AND THE EFFECT OF THE SIDNIE SYSTEM OVERALL ON THE CHANGE IN TEST SCORES, WITHIN EACH EXPERIMENT.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Criteria</th>
<th>Condition</th>
<th>N</th>
<th>Pre-test Scores Mean SD</th>
<th>Post-test Scores Mean SD</th>
<th>Condition Effects F-test p</th>
<th>Overall SIDNIE Effect F-test p Interaction?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction fidelity</td>
<td>Unbiased</td>
<td>High fidelity</td>
<td>8</td>
<td>97.30 3.44</td>
<td>100.00 0.00</td>
<td>F(1,16)=0.62 0.44</td>
<td>F(1,16)=6.77 &lt; 0.01 no</td>
</tr>
<tr>
<td></td>
<td>Low fidelity</td>
<td>10</td>
<td>98.67 2.81</td>
<td>100.00 0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age appropriate</td>
<td>High fidelity</td>
<td>9</td>
<td>81.67 20.68</td>
<td>95.31 9.33</td>
<td>F(1,17)=0.55 0.47</td>
<td>F(1,17)=13.47 &lt; 0.01 no</td>
</tr>
<tr>
<td></td>
<td>Low fidelity</td>
<td>10</td>
<td>74.00 21.42</td>
<td>94.50 6.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual fidelity</td>
<td>Unbiased</td>
<td>High fidelity</td>
<td>25</td>
<td>99.73 1.33</td>
<td>100.00 0.00</td>
<td>F(1,49)=3.10 0.08</td>
<td>F(1,49)&lt;0.01 0.99 no</td>
</tr>
<tr>
<td></td>
<td>Low fidelity</td>
<td>26</td>
<td>99.42 2.05</td>
<td>99.17 2.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age appropriate</td>
<td>High fidelity</td>
<td>25</td>
<td>68.89 18.84</td>
<td>94.31 13.40</td>
<td>F(1,48)=1.23 0.27</td>
<td>F(1,48)&lt;0.01 0.99 no</td>
</tr>
<tr>
<td></td>
<td>Low fidelity</td>
<td>25</td>
<td>65.73 15.02</td>
<td>90.24 12.30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE II. STATISTICS FOR PRE- AND POST-TEST SCORES FOR QUESTIONS WRITTEN BY STUDENTS AND SCORED BY EXPERTS. IN THIS TABLE, THE SCORES ARE CALCULATED WITHIN EACH EXPERIMENT, AND ACROSS BOTH EXPERIMENTS. THE STATISTICAL TESTS MEASURE THE INFLUENCE OF THE EXPERIMENT ON CHANGE IN TEST SCORES, AND THE INFLUENCE OF THE SIDNIE SYSTEM OVERALL ON THE CHANGE IN TEST SCORES.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Criteria</th>
<th>Pre-test Mean SD</th>
<th>Post-test Mean SD</th>
<th>Experiment Effect F-test p</th>
<th>SIDNIE Effect F-test p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction</td>
<td>Unbiased</td>
<td>98.15 1.40</td>
<td>100.00 0.00</td>
<td>2.39 &lt; 0.01</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Age appropriate</td>
<td>77.63 20.88</td>
<td>94.88 7.93</td>
<td>4.20 &lt; 0.01</td>
<td></td>
</tr>
<tr>
<td>Visual</td>
<td>Unbiased</td>
<td>99.58 1.73</td>
<td>99.58 1.73</td>
<td>0.16 &lt; 0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age appropriate</td>
<td>67.31 16.93</td>
<td>92.28 12.89</td>
<td>3.54 &lt; 0.05</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>Unbiased</td>
<td>99.20 2.23</td>
<td>99.69 1.49</td>
<td>0.64 &lt; 0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age appropriate</td>
<td>70.15 18.53</td>
<td>92.99 11.74</td>
<td>0.64 &lt; 0.05</td>
<td></td>
</tr>
</tbody>
</table>

V. LEARNING OUTCOMES

Using our pre- and post-questionnaire designed to measure learning outcomes, we compared questionnaire results over time and across conditions within each experiment. We also used the same questionnaire across both experiments so that we could compare outcomes across experiments to determine whether one experiment yielded better learning outcomes than the other. To analyze this data, we performed MANOVA tests using the experiment and the condition as independent variables and the questionnaire scoring as a repeated measure and the dependent variable.

A. Questions Written by Participants

To obtain scores for participant-written questions, three scorers evaluated each question on the basis of whether it was age appropriate and unbiased. We then averaged the scorers votes, yielding a percentage value for age appropriateness and unbiasedness.

Several participants were excluded due to missing questionnaire data (for example, they forgot to record their participant ID or submitted the form before completing it). This yielded 8 participants in the high fidelity condition and 10 participants in the low fidelity condition for the interaction fidelity study, and 25 participants in the high fidelity condition and 26 participants in the low fidelity condition for the visual fidelity study. For this analysis, we also excluded one outlier in the interaction fidelity experiment when scoring for bias, and one outlier in the visual fidelity experiment when scoring for age appropriateness.

For scoring on unbiasedness for each experiment, we encountered a ceiling effect, where every participant scored 91.67% or above on the pretest, leaving little room for improvement. In fact, all participants in the interaction fidelity experiment scored 100% on their post-test for unbiasedness, while all participants in the visual fidelity experiment scored 99% on their post-test for unbiasedness as well. This may indicate that we should choose a more difficult question criteria to better evaluate learning outcomes. In scoring for age appropriateness, scores were lower and had a wider range, with pretest scores ranging from 33.33% to 100.00%. Within both the visual and interaction fidelity studies, there was no significant difference between the low and high fidelity conditions for the change in scores between pretest and post-test.

We also tested within subjects effects to determine if in either experiment the interaction with SIDNIE served to significantly improve post-test scores (regardless of condition). Post-test scores improved significantly for both unbiasedness and age appropriateness in the interaction fidelity study, and improved significantly for age appropriateness in the visual fidelity study. In these tests, there was no interaction effect between the participant’s condition and the experiment.

We also wanted to investigate whether the two experimental groups performed differently on the pre- and post-tests, since they came from different demographic groups. There was a significant difference in the change of questionnaire scores for
age appropriateness, where participants in the visual fidelity experiment scored lower than the participants in the interaction fidelity experiment. Finally, we wanted to determine, regardless of experiment and condition, whether scores significantly changed after interaction with the SIDNIE system. For both age appropriateness and unbiasedness, participants showed a significant increase in post-test scores. For unbiasedness, there was a significant interaction effect between the experiment and the progress over time, where participants in the visual fidelity experiment did not improve their average score over time for age appropriateness (with average pre- and post-test scores being above 99%, showing a strong ceiling effect), while participants in the interaction fidelity study did show score improvement between the pre- and post-tests (from an average of approximately 98% to 100%). For detailed statistics refer to Tables I and II.

### B. Questions Scored by Participants

Participants scored five given questions for their age appropriateness and unbiasedness. We calculated percentage scores by totaling the number of correct scorings within each criteria, then dividing that number by five to yield a percentage score.

We excluded several participants due to missing questionnaire data (for example, the participant forgot to record their ID on the questionnaire, or submitted the form before filling in any answers). For the interaction fidelity experiment, this yielded 7 participants in the high fidelity condition and 8 participants in the low fidelity condition. For the visual fidelity experiment, this yielded 27 participants in the high fidelity condition and 25 participants in the low fidelity condition.

For this task, scores were lower and had a wider range for both pre-test and post-test, as could be expected for a task evaluating a higher level of learning. Both pre-test and post-test scores for age appropriateness ranged from 20% to 100%, while pre- and post-test scores for unbiasedness ranged from 40% to 100%.

Within both the visual and interaction fidelity experiments, there was no significant difference in the change of average pre- and post-test scores due to the participant’s condition. When we analyzed the scores within experiments and within subjects (regardless of condition), we found that the only significant change in average pre- and post-test score was for participants in the visual fidelity experiment for the criteria of age appropriateness.

Similarly to the analysis for questions written by participants, we also analyzed scores across experiments to determine whether the experimental groups performed differently on the pre- and post-tests. There was no significant difference between the two groups. Lastly, we wanted to determine, regardless of experiment and condition, whether scores significantly changed after interaction with SIDNIE. There was a significant difference in the average score for age appropriateness between the pre-test and post-test. There was no significant difference in the average score for unbiasedness; however, there was an interaction effect that showed that in the interaction fidelity experiment, scores decreased from pre-test to post-test, while in the visual fidelity experiment, scores increased from pre-test to post-test. For detailed statistics refer to Tables III and IV.

### VI. Qualitative Feedback

Overall, our system scored high on the System Usability Scale, with individual overall scores ranging from 65% to 98.33%, with an average score of 85% (sd=0.08) for the visual fidelity study and an average score of 84.3% (sd=0.09) for the interaction fidelity study. There were no significant differences in overall SUS score due to the participant’s condition in either experiment. When we analyzed the SUS scores question by question, we only found one significant difference. In the visual fidelity experiment, in response to the question ‘I found the system very cumbersome for use’, participants who were in the high fidelity (animated) condition found the system less cumbersome than those in the low fidelity (screenshot) condition (F(1,52)=4.80, p < 0.05). A chart showing SUS results by experiment and condition is provided in Figure 3.

#### A. Co-presence

We also asked participants questions about their sense of co-presence with the characters in the virtual environment.

---

**TABLE III. Statistics for pre- and post-test scores for questions students scored for age appropriateness and unbiasedness.** In this table, scores are calculated within each condition and within each experiment. The statistical tests measure the effect of the condition on the change in test scores, and the effect of the SIDNIE system overall on the change in test scores, within each experiment.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Criteria</th>
<th>Condition</th>
<th>N</th>
<th>Pre-test Scores</th>
<th>Post-test Scores</th>
<th>Condition Effects</th>
<th>Overall SIDNIE Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Interaction fidelity</td>
<td>Unbiased</td>
<td>High fidelity</td>
<td>7</td>
<td>88.57</td>
<td>15.74</td>
<td>88.57</td>
<td>15.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low fidelity</td>
<td>8</td>
<td>92.50</td>
<td>10.35</td>
<td>75.00</td>
<td>17.73</td>
</tr>
<tr>
<td></td>
<td>Age appropriate</td>
<td>High fidelity</td>
<td>7</td>
<td>80.00</td>
<td>23.09</td>
<td>88.57</td>
<td>15.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low fidelity</td>
<td>8</td>
<td>72.50</td>
<td>23.75</td>
<td>87.50</td>
<td>10.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High fidelity</td>
<td>27</td>
<td>86.67</td>
<td>15.69</td>
<td>94.07</td>
<td>14.48</td>
</tr>
<tr>
<td>Visual fidelity</td>
<td>Unbiased</td>
<td>High fidelity</td>
<td>27</td>
<td>84.80</td>
<td>17.59</td>
<td>87.20</td>
<td>18.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low fidelity</td>
<td>25</td>
<td>73.33</td>
<td>24.18</td>
<td>86.67</td>
<td>20.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low fidelity</td>
<td>25</td>
<td>72.00</td>
<td>24.49</td>
<td>86.40</td>
<td>19.77</td>
</tr>
</tbody>
</table>

---

**TABLE IV. Statistics for pre- and post-test scores for questions students scored for age appropriateness and unbiasedness.** In this table, the scores are calculated within each experiment, and across both experiments. The statistical tests measure the influence of the experiment on change in test scores, and the influence of the SIDNIE system overall on the change in test scores.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Criteria</th>
<th>Condition</th>
<th>Pre-test</th>
<th>Post-test</th>
<th>Experiment Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>F(1,165) p</td>
</tr>
<tr>
<td>Interaction</td>
<td>Unbiased</td>
<td>High fidelity</td>
<td>90.67</td>
<td>12.80</td>
<td>81.33 17.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low fidelity</td>
<td>76.00</td>
<td>22.93</td>
<td>88.00 12.65</td>
</tr>
<tr>
<td></td>
<td>Age appropriate</td>
<td>High fidelity</td>
<td>85.77</td>
<td>16.49</td>
<td>90.77 16.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low fidelity</td>
<td>72.69</td>
<td>24.10</td>
<td>86.54 19.69</td>
</tr>
<tr>
<td>Visual</td>
<td>Unbiased</td>
<td>High fidelity</td>
<td>86.87</td>
<td>15.78</td>
<td>88.66 17.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low fidelity</td>
<td>73.43</td>
<td>23.71</td>
<td>86.87 18.27</td>
</tr>
<tr>
<td>Overall</td>
<td>Unbiased</td>
<td>High fidelity</td>
<td>90.67</td>
<td>12.80</td>
<td>81.33 17.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low fidelity</td>
<td>76.00</td>
<td>22.93</td>
<td>88.00 12.65</td>
</tr>
<tr>
<td></td>
<td>Age appropriate</td>
<td>High fidelity</td>
<td>85.77</td>
<td>16.49</td>
<td>90.77 16.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low fidelity</td>
<td>72.69</td>
<td>24.10</td>
<td>86.54 19.69</td>
</tr>
</tbody>
</table>

---

**TABLE V.** Statistical test results by experiment and condition is provided in Figure 3.
We found no significant differences across experiments, or between the conditions for the co-presence questionnaire in the visual fidelity experiment. There was, however, a significant difference between the average scores for three questions in our co-presence questionnaire according to the participant’s condition within the interaction fidelity experiment. For the prompts “To what extent, if at all, did the virtual patients hinder you from carrying out the task?” and “To what extent, if at all, were there times during which the computer interface seemed to vanish, and you were working directly with the virtual patients?”, participants in the high fidelity (simulated voice recognition) condition agreed more strongly with the prompt than those in the low fidelity (clicking questions) condition. In response to the question “How realistic were the virtual patients (for example, how they looked, moved, spoke and interacted with you)?”, participants in the high fidelity (simulated voice recognition) condition rated the character realism higher than those in the high fidelity (simulated voice recognition) condition (all p’s <0.05). Although the differences were not significant, within the interaction fidelity experiment, the mean score for participants in the high fidelity condition was higher than the mean score for the participants in the low fidelity question in all but three questions: “To what extent did you feel embarrassed with respect to what you believed the virtual patients might be thinking of you?”, “To what extent, if at all, was working with this system like working with patients in the real world?”, and “How realistic were the virtual patients (for example, how they looked, moved, spoke and interacted with you)?”. Based on this observation along with the significant differences for that experiment, it seems that using the higher fidelity interaction metaphor may have made the participant have higher expectations for other aspects of the interaction, which the simulation fell short of. Results for the visual fidelity experiment were more mixed, with six out of the fourteen questions receiving higher mean scores in the low fidelity condition. This may indicate a similar trend where low fidelity characters lead to low expectations of realism and allow the participant to subconsciously fill the gaps in realism instead of drawing attention to unrealistic elements in the characters’ behavior.

Overall, our co-presence scores were low, as indicated by the count of the number of questions which received a mean score of 4 out of 7: 4 questions out of 14 questions, for both experiments. The highest rated questions in both experiments for co-presence were “To what extent, if at all, did the virtual patients help you in carrying out the task?” and “To what extent, if at all, were you worried about what you believed the virtual patients might be thinking of you?”. While in the visual fidelity experiment, the similar question, “To what extent did you feel embarrassed with respect to what you believed the virtual patients might be thinking of you?” The low scores on these questions may show that the participants were confident in their abilities instead of indicating a low sense of co-presence with the virtual patients. See Table V for mean scores for each of the co-presence questions divided by experiment and condition.

**B. Debriefing Interview**

Through a debriefing interview, we asked participants open-ended questions on their opinions of the system. We tailored each interview to the experiment’s focus and also asked questions about the overall system in both experiments.

1) **Interaction Fidelity:** When asked if they could choose to interact with the system in any way, six of the 22 participants in the interaction fidelity study said they would choose speech interaction. Two participants preferred to click the questions, while the remaining participants were neutral or listed several interaction options they considered equivalent. Participants seemed to have some level of anxiety about speaking to the
system—six participants made comments about being self-conscious when speaking, saying that talking was “intimidating”, “awkward when you are in a room with other people”, or that they “would feel like an idiot talking to the computer”. However, participants also expressed that they thought that speaking seemed more realistic and useful for practice than clicking on the questions, commenting that “[talking] might be good...it’s harder for me to initiate the talking, so the practice of speaking might be good because with a real patient you can’t click a button”, and that talking was “much more challenging and I felt I needed to be more professional and integrate speech”.

All but three participants in the interaction fidelity study liked the guidance aspects of the system, such as the question-by-question feedback and being able to select from four options of questions to ask. Participants commented that “the choices give a good starting point if someone has no experience at all” and that “starting out I would like having the choices so I would know exactly what I should be asking”. However, thirteen of the participants recognized the limitations of being able to select from the four questions for learning purposes, and suggested that the interaction should progress from selecting questions from options to asking free-form questions in natural language as students become more experienced, commenting that “as I learn how to interview a patient it’d be nice to have some freedom” and “as I progress further in nursing I would rather make my own [questions]”.

Because speech recognition adds an additional level of complexity on top of natural language processing for asking free-form questions, we asked the participants whether they would like to type the question. Fourteen of the participants said that typing in their questions would be a good interaction technique. Some participants said they would prefer to type because they would be less self-conscious about their performance when speaking, or that with typing they would have more time to think about their question and make corrections before submitting it to the patients. However, eight participants still commented that speaking would be more useful than typing, saying that “practicing speech with the patient is more valuable than you might think” and “you have to get comfortable with saying those questions”.

2) Visual Fidelity: When we asked participants if there was anything about the picture or animations (depending on their experimental condition) that they did not like, participants who had the animated characters often commented on the lack of realism in the characters but seemed overall satisfied with their appearance, while participants with the static image commented on the lack of animation. Some participants in the static image condition were clearly displeased with the lack of animations, making comments such as “I didn’t like that they didn’t move”, “They didn’t move at all and weren’t breathing or blinking”, and “I didn’t like how they were so still”. Curiously, however, at least two participants did not know that they were in the static image condition, saying, “Were they really still the whole time? I didn’t notice that. Guess I didn’t really pay attention to them” and “I could have sworn I saw the mom move”. Participants who thought that the patients were realistic often qualified their assessment, saying that it “looked like a video game but you could tell that they were humans”, “they looked pretty realistic for a computer person”, “they were realistic but still obviously computer animated”, and “they were pretty realistic compared to other virtual things”.

When asked if there was anything visually they would like to change, most participants in the static image condition said that they wanted the patients to be animated, while participants in the animation condition mostly wanted more “passive” animations, such as interaction between the mother and the child or more fidgety behavior for the child. When asked about the mother’s animations specifically, 27 participants changed the topic and commented on the mother’s answers, commenting...
that her text-to-speech voice sounded computerized or angry, or that her responses to questions sounded unfriendly or short. Seven participants commented that the mother was not as compassionate towards her daughter as they expected, both in terms of behavior and speech, saying that the mother and child were not as physically close as they would have expected as well as that the mother did not elaborate much on her answers to medical questions. When asked about the child’s animations specifically, 25 participants said the child was unrealistic due to her lack of movement (“the child just sat there”) or lack of contribution to the interview (“she didn’t really talk much at all”). Participants seemed to reflect a variety of ideas of what age appropriate for a five-year old, with some participants saying “She didn’t really talk, but she’s 5 though, so it’s okay” and “I would have imagined that the child would be more verbal since they’re in pain”.

To gauge how much participants looked at the patients at all, we asked participants to report how often they looked at the patients. Twenty nine participants looked at the patient all, we asked participants to report how often they looked at the patients. Twelve participants reported looking at the patients every two or three questions. Ten participants reported looking at the patients only every once in a while, with seven of those participants explicitly stating that they stopped looking at the patients because they were not animated.

Overall Feedback. Participants in both experiments made comments about the system’s lack of visual realism. Forty-one participants commented that the characters were unrealistic in some aspect, citing reasons such as the “robotic” text to speech voices, lack of smooth animations, and lack of interaction between the mother and daughter. In particular, the participants felt the child was unrealistic in terms of her behavior, since she showed “no physical signs of illness” and since the child answered few questions in comparison to the mother. One participant who worked with children on a regular basis said that she “would have expected [the child] to say more things since she was nearly six; [she] probably would have said more about how it felt and where it hurts than just ‘it hurts’. [I] would have expected that from a 3 year old”. However, the participant also acknowledged that the interviewing a patient “requires you to know the family or seeing how it goes...maybe you ask questions, see if the child is able to give useful information, if not, you’ll have to interact more with the parent,” so a wide range of behavioral characteristics could be appropriate. Comments about the characters’ physical appearance included, “they looked like they were related” and “the mom was young...the child’s skin tone was too dark for her hair”. Several participants also commented that the virtual environment was not realistic, saying that the murals on the walls were too busy or the lighting was too dim.

Overall, participants seemed to enjoy their interaction with the system, commenting that “I just wanted to see how they were reacting to [me]” and “I enjoyed this”. All but one of the participants in the visual fidelity experiment said that they thought they and other students would use this system if it were available to them, although the answer was often qualified by saying that they would want more scenarios to practice with to prevent it from becoming repetitive. Participants in both experiments commented that they thought it would help them be less nervous when interacting with actual patients, saying that “it makes it less awkward than if you were sitting in the room or asking the wrong questions”, “it would make me more comfortable before going into the [doctor’s] office and hospital”, and “you can go through situations that you wouldn’t be able to in real life; you get to practice”.

VII. DISCUSSION

Our primary hypotheses for both the interaction and visual fidelity experiment were confirmed to be partially true. Participants preferred the high fidelity conditions, but the learning outcomes in either condition were positive. The students produced high scores in both the low and high fidelity conditions. Even though there was not a significant difference between conditions in the learning outcomes, when it came to unbiasedness we noticed that the participants scored above a 91% in both the pre- and post-tests, which may be attributed to our question criteria being too simple. We also noticed that regardless of experiment and condition, participants showed a significant increase in post-test scores in both unbiasedness and age appropriateness. We can conclude that even if we do not include the animations and speech recognition in our system, students will still achieve the proper learning outcomes.

Our system produced high scores on the System Usability Scale for both experiments, showing that SIDNIE is not difficult to use. An interesting result from our co-presence questionnaire in the interaction fidelity study revealed that students scored the low fidelity condition (clicking) as being more realistic than the high fidelity condition (speaking). As noted in the results above, this may be because participants have lower expectations when they are given a low fidelity system and high expectations when it comes to the high fidelity system.

In the interaction fidelity experiment, most participants expressed that they would prefer to talk to the system and that speaking to the system is more realistic and would be useful for practicing. We also asked the participants of the interaction fidelity experiment whether they would like to type out their questions instead. Most participants commented that typing out their questions would give them the option to ask a larger variety of questions, but they also understood that practicing speaking out loud was very important.

In the visual fidelity experiment, most participants seemed to be satisfied with the overall appearance of the characters and the environment, but the participants who had the high fidelity condition commented on the lack of realism of the animations, while the participants in the low fidelity condition asked for the characters to be animated. Even though the high fidelity participants commented on the lack of realism of the animations they still preferred to include them. A few interesting comments from the debriefing interview brought to our attention the fact that some of the participants in the low fidelity condition thought they had animated characters. Some of the suggestion given in this experiment were to add more passive animations, particularly for the child, and more animations related to the interaction between the mother and the child.

VIII. FUTURE WORK

Overall, the results of the two fidelity experiments were positive and we gained some valuable suggestions for improve-
ment from the participants. We would like to continue to evaluate learning outcomes, specifically with respect to SIDNIE’s scaffolding capabilities in combination with various levels of behavioral and interaction fidelity. However, due to the ceiling effects shown in this experiment, we need to develop a greater variety of scenarios and scoring criteria. Currently, one of the largest obstacles with creating virtual patient systems in the nursing field is the creation and generation of the scenarios needed by the system. It is a large undertaking that requires a significant amount of time and expertise. To reduce the workload on both nurses and virtual world developers, we are creating a tool for nursing faculty to be able to input their own scenarios with categories and grading criteria, with some automated generation of questions, answers, and animations based on minimal nurse input. Additionally, we aim to build a database of questions and answers that may be reused for multiple scenarios to further reduce workload.

The results of our studies also show the importance of more realistic animations and adding speech recognition to the system. Participants also suggested changing the voices to better represent empathy and concern. After we make these suggested changes we would like to dive deeper into the different aspects of each scaffolded level and to ensure that our levels are in concurrence with Bloom’s Taxonomy [21]. We would also like to test the learning outcomes from using the scaffolded system to ensure that each level is taking the student in the correct direction.

ACKNOWLEDGMENT

This research was supported, in part, by the NSF Graduate Research Fellowship Program (fellow numbers 2009080400 and 2011095211), an Interdisciplinary Research Innovations Grant from the College of Health Education and Human Development at Clemson University, and a grant from the Agency for Healthcare, Research, and Quality (RO3#HS020233-01).

REFERENCES


