5-2015

Implementing and Evaluating a Scenario Builder Tool for Pediatric Virtual Patients

Lauren Elizabeth Cairco Dukes
Clemson University

Implementing and Evaluating a Scenario Builder Tool for Pediatric Virtual Patients

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Computer Science

---

by
Lauren Elizabeth Cairco Dukes
May 2015

Accepted by:
Dr. Larry F. Hodges, Committee Chair
Dr. Sabarish Babu
Dr. Donald House
Dr. Timothy Davis

Abstract

Baccalaureate nursing students have few opportunities to practice patient interaction before they reach their clinical experiences. Traditional practice opportunities include roleplay and interviews with paid actors (called standardized patients). Unfortunately, neither of these methods realistically simulates many of the patient interactions that nurses will encounter on a daily basis.

Virtual patients are computer simulations that behave in the same way that an actual patient would in a medical context. Since these characters are simulated, they can provide realistic yet repetitive practice in patient interaction since they can represent a wide range of patients and each scenario can be practiced until the student achieves competency. However, the development costs for virtual patients are high, since creation of a single scenario may take up to nine months.

In this work, we present a virtual patient platform that reduces development costs. The SIDNIE (Scaffolded Interviews Developed by Nurses in Education) system can adapt a single scenario to multiple levels of learners and supports the selection of multiple learning goals. We have shown that SIDNIE is effective for learning [Dukes et al., 2013]. We designed and evaluated a scenario-builder tool that enables nursing faculty to create their own scenarios for SIDNIE, without the aid of a computer scientist. Additionally, we showed that scenarios created using this system could be effective for teaching nursing students verbal communication skills by conducting a user study with freshman nursing students.

Dedication

To my parents, David and Wendy Cairco, who told me to climb the ladder as high as I could go, but to not forget them when I got to the top! I will keep climbing, thanks to your unending prayers and support.

Acknowledgments

Thank you to the NSF Graduate Research Fellowship Program (fellow ID: 2009080400) for funding three years of my graduate research. My work was also funded in part by 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).

Thank you to my advisor, Dr. Larry F. Hodges for taking the chance on me nearly ten years ago when I applied for an undergraduate research position! It has been a privilege and a joy to learn from you. You have been the best advisor I could have imagined. Thanks to Mrs. Elizabeth Hodges, too, for adopting me into the family. I respect and trust both of you so much.

I would like to thank Winthrop University for providing me with an excellent undergraduate education that was foundational to all my graduate work. Dr. Lynn DeNoia, Dr. Will Thacker, and Dr. Anne Olsen have been excellent advisors through undergraduate and beyond.

Thank you also to Dr. Scott McCrickard, Dr. Shahtab Wahid, and Dr. Stacy Branham for your mentorship during my time at Virginia Tech. Perhaps you can see your fingerprints on the inspiration for the scenario builder tool. To Dr. David Krum, Dr. Adam Jones, and Thai Phan: thank you for making my time at the University of Southern California Institute for Creative Technologies wonderful.

To the members of the Virtual Environments Group, past and present: I owe each of you a great deal. It is an honor to be named among you. I have received inspiration and encouragment from so many of you through the years as both coworkers and mentors.

Jerome McClendon has been a wonderful encourager and friend for the entire time I've been here–from him, I learned that "nothing comes to a sleeper but a dream"–and we have both sacrificed much sleep to achieve our goals! Toni Pence has been a friend and coworker for the past decade as

we have traveled through all of our college years together. It would have been much less fun without you, Toni. Thank you to Elham Ebrahimi, Tania Roy, and Yvon Feaster for their friendship, advice, help, and sense of humor.

My work would have been literally impossible without the patient help of Dr. Nancy Meehan. Nancy, thank you for going above and beyond to make sure that all this happened. More importantly, thank you for being my best cheerleader and a great role model. Dr. Arlene Johnson was also a great help in providing crucial nursing knowledge.

All the faculty and staff in the School of Computing have been wonderfully supportive, cooperative, and encouraging. Thank you, Dr. Brian Dean, for being a patient teacher, a good listener, and for all the chocolate chip cookies. Thank you to Nell Kennedy for keeping all my computers and software working and for being an encouragement through many difficult days.

Some dear friends have provided support and inspiration throughout the years: Rachel Grotheer, Andrew Smith, Bethany Bush, Lea Queener, Alex Godwin, and Kaci Kawakami. I am especially grateful for my church family at Crosspoint and for their relentless care. I have felt the care and support of hundreds of others throughout this journey–too many to possibly list individually.

Thank you to my parents, David and Wendy Cairco, for your constant encouragement and your belief in me. Thanks to my sister Victoria for being a good friend throughout the years. Thanks to my sister Briana for the cookies and tater tots left at my desk many days, and for the free meals from the Clemson cafeterias!

I am grateful for the support and kindness of my husband, Patrick, throughout this process. I could not have survived without his patient care.

Most of all, I "praise God from whom all blessings flow". My time here has been a true blessing. May Christ be honored by this work.

Table of Contents

Title Page . . . . . **i**

Abstract . . . . . **ii**

Dedication . . . . . **iii**

Acknowledgments . . . . . **iv**

List of Tables . . . . . **viii**

List of Figures . . . . . **x**

1 Introduction **1**

2 Background and Related Work **4**
2.1 Simulation Learning in Healthcare 4
2.2 Assessing Learning in Virtual Patient Scenarios 6
2.3 Scaffolding for Learning 7
2.4 Authoring Virtual Patient Scenarios 9

3 Initial Virtual Patient Prototype: The Virtual Pediatric Patient System **13**
3.1 Study Description 16
3.2 Results and Discussion 17
3.3 Lessons Learned 20

4 SIDNIE: Scaffolded Interviews Developed by Nurses In Education **21**
4.1 Scenario Development 22
4.2 Sidnie 22
4.3 Tutorial 23
4.4 Electronic Health Record 23
4.5 Pediatric Patient Interview 23
4.6 Feedback Mechanisms 26
4.7 Scaffolding Levels 27
4.8 Implementation Details 29

5 Evaluation and Refinement of SIDNIE **32**
5.1 The Effects of Scaffolding on Learning Outcomes 32
5.2 The Effects of Visual and Interaction Fidelity on Learning 38
5.3 Modifications to SIDNIE 56

6 Participatory Design of the Scenario Builder Tool **59**
6.1 Early Prototype 59

6.2 Participants and Experimental Design . . . 65
6.3 Session One: Introduction . . . 68
6.4 Session Two: Initial Prototype . . . 68
6.5 Session Three: Refined Prototype . . . 78
6.6 Session Four: Final Prototype . . . 82
6.7 Discussion . . . 95
6.8 Recommendations for Future Participatory Designs . . . 95

7 Usability Evaluation of the Scenario Builder Tool . . . 97
7.1 Changes to Scenario Wizard for Usability Study . . . 97
7.2 Experimental Procedure . . . 106
7.3 Results . . . 108
7.4 Discussion . . . 117
7.5 Recommendations for Designing Content Creation Interfaces . . . 119

8 Evaluation of Learning Scaffolding . . . 121
8.1 Experimental Procedure . . . 122
8.2 Results . . . 123
8.3 Discussion . . . 141
8.4 Comparing Results with Other Studies . . . 144
8.5 Recommendations for Other Applications . . . 147

9 Conclusion and Future Directions . . . 149
9.1 Future Directions . . . 150

Appendices . . . 153
A List of Related Publications . . . 154
B Informed Consent: Usability Evaluation . . . 155
C Demographic Questionnaire: Usability Evaluation . . . 158
D Nursing Interviewing Skills Rubric . . . 159
E Postquestionnaire: Participatory Design . . . 162
F Case Study: Usability Evaluation . . . 164
G Post-Task Questionnaire: Usability Evaluation . . . 165
H Post-Interview: Usability Evaluation . . . 166
I System Usability Scale . . . 167
J Informed Consent: Scaffolding Evaluation . . . 168
K Demographic Questionnaire: Scaffolding Evaluation . . . 171
L New General Self-Efficacy Scale . . . 172
M Learning Outcomes: Scaffolding Evaluation . . . 173
N Presence and Copresence Questionnaires . . . 174
O Post-Interview: Scaffolding Evaluation . . . 176

Bibliography . . . 177

List of Tables

3.1 Categorized questions-answer pairs by participant. Reported values are percentages of the total questions the participant asked. 19

5.1 Demographic information for the study participants (1= no experience, 7= very experienced) 33
5.2 Percentage means and standard deviation for each scaffolding level. 35
5.3 Percentage means and standard deviation for each scaffolding level. 36
5.4 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. There were no interaction effects in this test. 42
5.5 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. 42
5.6 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. There were no interacton effects in these tests. 46
5.7 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. 46
5.8 Mean ratings on a scale of 1 to 7 for the co-presence questionnaire. Scores for questions with a * were reversed, so that higher scores always indicate greater co-presence. H stands for high fidelity condition, while L stands for low fidelity condition. 51

7.1 Demographic information for the seven participants in the usability study. In the "Role" row, F stands for faculty member, while S stands for student. 108
7.2 This table shows the amount of time faculty participants took to complete each step in the scenario creation process in minutes and seconds. 109
7.3 Completion times for the final two tasks for faculty members. The third column is a projected completion time for the task of creating a new criteria, since the task was only completed in part in the usability study due to time constraints. 110
7.4 Difficulty rankings for each task completed by faculty participants, where 1 is the least difficult, and 5 is the most difficult. 110
7.5 Faculty participants reported which task was the most and least difficult. 111

7.6 Responses for the system usability scale by participant. 111

8.1 This table shows the average amount of experience participants reported in several categories, where 1 was labeled as “none” and 4 was labeled as “a lot”. 124

List of Figures

3.1 A student interacting with the Virtual Pediatric Patient System. 14

4.1 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. 24
4.2 When summative feedback is provided, at the end of the scenario, a chart is displayed with the score for each question the user asked. 25

5.1 Student interacting with SIDNIE via speech recognition. 39
5.2 Results of the SUS questionnaire separated out by condition and experiment. Scores for questions with a * have been reversed so that in this chart, higher scores always represent better usability. 48
5.3 Screenshot of the improved doctor's office with SIDNIE. 57

6.1 Three examples of five year old Caucasian females generated by the first iteration of the character building scripts. 61
6.2 Three examples of ten year old African males generated by the first iteration of the character building scripts. 61
6.3 Nurse educators will first be encouraged to select a character from a library, encouraging reuse. 62
6.4 If there is no suitable character, the nurse may edit an existing character or create a new one. 63
6.5 The nurse educator fills out the information for the patient's electronic health record. 63
6.6 The nurse educator checks off the symptoms that the patient should present. This is used to automatically select some questions and answers for the scenario, and to generate animations. 64
6.7 The nurse educator selects the learning goals for his or her scenario. 64
6.8 The nurse educator may add new learning goals not already present in the system. . 65
6.9 The nurse educator can review the questions and answers automatically selected based on her input into the system. 66
6.10 If the nurse educator wants to add a question, he or she is first directed to the question library to see if there are any suitable questions already present. 66
6.11 The nurse educator may add a new question or edit an existing question to better suit the simulation. 67
6.12 After the nurse completes scenario generation, he or she can preview the simulation in each scaffolded learning level before packaging it for student use. 67
6.13 This figure shows the prototype running inside the Blend interface. The drawing and commenting tools are located at the bottom left corner of the screen. 69
6.14 Screenshot of the lesson plan screen for the initial prototype. 70
6.15 Screenshot of the previewer screen for the initial prototype. 71
6.16 Screenshot of the patient selection screen for the initial prototype. 72

6.17 Screenshot of the chief complaint screen for the initial prototype. 73
6.18 Screenshot of the history of present illness screen for the initial prototype. Each component of the history of present illness is represented by an expander with several options below. 74
6.19 Screenshot of the review of systems screen for the initial prototype. Each body system is represented by an expander with several related symptoms that can be selected. . 75
6.20 Screenshot of the electronic medical record screen for the initial prototype. Each section of the electronic medical record is represented by a tab. Many fields in the electronic medical record were automatically populated according to the patient's demographics. 76
6.21 This figure shows the history of present illness screen in the second prototype. On the left hand side of the screen, there are buttons to return to previously completed sections, as well as a listing of the sections to come. 79
6.22 A screenshot of the preview screen for the questions automatically generated using the information in the scenario builder. 80
6.23 A screenshot of the lesson plan in the new Modern UI styled interface. 82
6.24 A screenshot of the top navigation bar with menus (large text) and submenus (smaller text). 84
6.25 A screenshot of the preview screen for the questions automatically generated using the information provided in the scenario builder. The interface is improved to make it easier to read. 84
6.26 This interface enables the user to select the demographic characteristics of the character they want to generate. 85
6.27 This interface enables the user to select more fine-grained appearance characteristics of their character, including hair and clothing colors, and hair and eyebrow styles. . 85
6.28 This dialogue enables the user to create a new chief complaint and to enter the typical review of systems for that chief complaint. 87
6.29 This is the electronic medical record's interface for the patient's allergies. To enter a new allergy, the user could click on the gray line under the headers and type the allergen and the corresponding reaction. 88
6.30 In the improved review of systems pages, symptoms are separated out into those typically relevant to the case ("Included Symptoms") and those typically irrelevant ("Excluded Symptoms"). 93
6.31 This is the electronic medical record's improved interface for the patient's allergies. Clicking on a the "Add Allergy" button launched a dialog box where the user could put in the allergen and reaction. 94

7.1 The updated customization interface shows thumbnail views of hair, eyebrows, and eyelashes with buttons to right and left to change their styles. 99
7.2 This figure shows three children created using the updated character generation software. 99
7.3 This figure shows three adults created using the updated character generation software.100
7.4 The updated previewer shows the selected patients in a doctor's office along with questions and answers. When the user clicks the button to preview a question, the patients respond with text-to-speech and lipsyncing. 101
7.5 To create a new chief complaint, for each system, the user first checks off the symptoms they consider to be relevant to the chief complaint. 102
7.6 After selecting the relevant symptoms, the user selects whether each relevant symptom is typically present or typically absent for their chief complaint. 103
7.7 After selecting the relevant symptoms, the user selects whether each relevant symptom is typically present or typically absent for their chief complaint. 103

7.8 When creating a new criteria, for each interview category the user is required to score all the existing questions as well as to write two new questions to contribute to the database. . . . 105
7.9 The user can select a patient from the library or choose to create their own character. 112

8.1 Boxplots for the pretest and posttest questions participants scored correctly for sensitivity. Whiskers show the minimum and maximum values. The median is indicated by a circle. . . . 125
8.2 Boxplots for the pretest and posttest questions participants scored correctly for appropriate language. Whiskers show the minimum and maximum values. The median is indicated by a circle. . . . 126
8.3 Questions that participants wrote in their prequestionnaire and postquestionnaire, scored as to whether they showed appropriate language. Error bars show one standard deviation. . . . 127
8.4 Boxplots for the pretest and posttest questions participants wrote, scored for appropriate language. Whiskers show the minimum and maximum values. The median is indicated by a circle. . . . 127
8.5 Questions that participants wrote in their prequestionnaire and postquestionnaire, scored as to whether they showed sensitivity. Error bars show one standard deviation. 128
8.6 Boxplots for the pretest and posttest questions participants wrote, scored for sensitivity. Whiskers show the minimum and maximum values. The median is indicated by a circle. . . . 129
8.7 Questions that participants scored correctly for sensitivity. Error bars show one standard deviation. . . . 130
8.8 Boxplots for the pretest, posttest, and e-mail questionnaires where participants scored questions for appropriate sensitivity. Whiskers show the minimum and maximum values. The median is indicated by a circle. . . . 130
8.9 Questions that participants scored correctly for language. Error bars show one standard deviation. . . . 131
8.10 Boxplots for the pretest, posttest, and e-mail questionnaires where participants scored questions for appropriate language. Whiskers show the minimum and maximum values. The median is indicated by a circle. . . . 131
8.11 Questions that participants wrote that had appropriate language. Error bars show one standard deviation. . . . 132
8.12 Boxplots for the pretest, posttest, and e-mail questionnaire scores where participants wrote questions with appropriate language. Whiskers show the minimum and maximum values. The median is indicated by a circle. . . . 133
8.13 Questions that participants wrote that had appropriate sensitivity. Error bars show one standard deviation. . . . 134
8.14 Boxplots for the pretest, posttest, and e-mail questionnaire scores where participants wrote questions with appropriate sensitivity. Whiskers show the minimum and maximum values. The median is indicated by a circle. . . . 134
8.15 Results of the self-efficacy questionnaire comparing pre-test and post-test scores. Error bars show one standard deviation in each direction. Questions labeled with a $*$ showed a significant change between pretest and posttest scores. . . . 135
8.16 This chart shows the means and standard deviations for the presence and copresence scores divided by condition. Questions with scoring reversed are indicated with a *. The error bars represent one standard deviation in each direction. . . . 138

Chapter 1

Introduction

During their baccalaureate education, nursing students have limited opportunities to practice patient interaction. Traditional training for patient interaction includes roleplay with peers, and practice interviews with paid actors. Unfortunately, neither of these methods adequately portrays the wide range of patients and medical conditions that nurses encounter in their clinical experiences. Since nursing students are mostly college aged females, many times roleplay fails to portray interactions between nurses and any other demographic group besides their peers. Additionally, there are populations for which it would be difficult to find standardized patients to portray–for example, children, people with disabilities, or people of differing ethnic groups.

Virtual patients are computer simulations that behave in the same way that an actual patient would in a medical context. There are many types of virtual patients, ranging from text-and-photo "interaction" to fully animated three dimensional virtual characters. Since these characters are simulated, they can provide realistic yet repetitive practice in patient interaction since they can represent a wide range of patients and each scenario can be practiced until the student achieves competency. Medical students have successfully used virtual patients to develop their clinical reasoning skills [Botezatu et al., 2010], communication skills [Lok, 2006, Stevens et al., 2006, Deladisma et al., 2009], interdisciplinary collaboration [Booth and McMullen-Fix, 2012, Reese et al., 2010] and patient interviewing skills [Johnsen et al., 2007].

Although text-and-photo based virtual patients are easier to produce since they can use standardized images and case studies, much of human communication is nonverbal [Thalmann, 2001] and so cannot be communicated through those channels. For animated virtual characters,

there are two options: recorded video clips that simulate interaction, and virtual characters. Virtual characters have the advantage of being more easily modified since recording new video with the same character after the intial production could be dfificult. However, there are two problems with existing virtual character and virtual patient platforms. First, *development costs* for virtual patients are high, since creation of a single scenario may take up to nine months. This process is time-consuming because it requires continual iteration between nurses, who create the content and determine the learning goals, and computer scientists, who implement the system. Often, each virtual patient is a completely new 3D model to be created and sometimes, a complete software platform change is necessary, depending on the simulation goals. The computer scientist is a bottleneck that all new content must pass through, and due to lack of domain knowledge in nursing, it often takes many iterations to come to a correct simulation. The second problem in virtual patient platforms is a *lack of extensibility*, since a scenario is typically only targeted towards a single set of learning goals and learners. After the cost of developing a scenario, it is often only good for one use per student, since there is no mechanism for quick adaptability to different learning goals. For nurses to be prepared for their clinical experiences, they must practice a wide range of scenarios, which current development techniques simply cannot provide in a timely manner.

In collaboration with the Clemson School of Nursing, we developed a virtual patient platform called SIDNIE (Scaffolded Interviews Developed by Nurses in Education). SIDNIE allows nursing students to interact with virtual patients while receiving guidance and feedback from a virtual nurse educator. The virtual patients respond verbally and nonverbally to nurse questions. SIDNIE's first scenario was effective for learning [Dukes et al., 2013, Pence et al., 2013], so the addition of more scenarios would extend its impact.

To address the cost of virtual patient authorship as well as to provide extensibility, we carried out the user-centered design and creation of a scenario-builder tool to enable nursing faculty to create their own scenarios for SIDNIE, without the aid of a computer scientist. The end product was a wizard-like tool that nursing faculty were able to use to select or create a virtual character customized to their desired demographic, then select the symptoms that their patient should exhibit and the scenario's desired learning goals. The scenario builder tool automatically selected questions and answers from a scored question bank, and adapted to the medical content of the scenario. The nurse educator could then customize or add questions and answers as desired, and then the scenario could be imported into the SIDNIE system. To encourage reuse, we created a centralized library of

characters, questions, answers, and scoring criteria that nursing educators may use and adapt. The scenario builder tool was designed using participatory design principles and was tested for usability in a formal study. Although other authorship tools have been proposed [Rossen et al., 2009, Ellaway, 2010, Medbiquitous, 2007], these systems only provide text-and-image based interaction, falling short of realistic simulation. This project is innovative in that it provides an efficient platform for nurse educators to generate complete, realistic scenarios that include believable conversation and embodied interaction. Additionally, we carried out further evaluation of SIDNIE's system to determine whether its scaffolded presentation yields better learning outcomes than a non-scaffolded presentation, using scenarios created by the scenario builder tool.

This thesis is organized in eight chapters and several appendices. Chapter 2 outlines related work about virtual patients and the virtual patient authorship process as well as scaffolding for improved learning outcomes. Chapter 3 describes our initial virtual patient system and the lessons we learned through its implementation. Chapter 4 describes the SIDNIE system, while Chapter 5 describes evaluations of its effectiveness for learning as well as improvements made to SIDNIE to better support learning and to integrate with the scenario generation tool.

Chapters 6 and 7 recount the design process for the scenario builder tool and its usability evaluation, while Chapter 8 shows the results of a study that evaluated SIDNIE's scaffolded learning capabilities using scenarios generated through the scenario builder tool.

Finally, the appendices provide the measures and other documents we used for our experimental evaluations.

Chapter 2

Background and Related Work

2.1 Simulation Learning in Healthcare

The Institute of Medicine has recommended the use of simulation training as a method to improve health care delivery [Stevens et al., 2006], and patient safety has been identified as one of the six competencies necessary to improve health care education [Johnsen et al., 2005]. Simulation training is an effective strategy to help promote safe clinical practices [Hubal et al., 2003] and impacts the development of self-efficacy and judgment skills for nurses that are essential to provide the safest and most effective care possible [Hockenberry et al., 2007]. 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 [Hockenberry et al., 2007]. 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 [Durham and Alden, 2008].

2.1.1 Current Simulation Practices for Nursing Communication

Mannequin simulation training is already used in nursing courses for many manual skills such as measuring vital signs and inserting IVs. However, designing effective simulations for patient communication training is significantly more difficult since it requires the simulation of a patient's cognitive state and abilities. Current simulation exercises in the Clemson University School of Nursing (and many other nursing schools) primarily involve students acting out written scenarios, where one student acts as the patient while the other acts as the nurse. Unfortunately, this simulation method cannot accurately portray the wide variety of patient scenarios that a nurse will soon encounter in his or her clinical experiences. Any physical condition or demographic that differs from the student's own cannot be observed, and the student can only answer questions that he or she already knows. Additionally, it is difficult for students to take these practice interviews seriously due to their lack of realism. Schools also may employ standardized patients (paid actors) as interview partners for nurses. Standardized patients may be advantageous over peer interview since they can represent a wider demographic. However, the time student nurses get to spend with standardized patients is extremely limited. Additionally, it is still difficult or unethical to represent many populations with a standardized patient such as individuals with certain disabilities or pediatric patients and their parents. Both of these simulation techniques are also difficult to replicate consistently for each nurse, and are difficult to score objectively. Also, there are few opportunities for real-time guidance and feedback from experienced practitioners or instructors during these interpersonal exchanges.

2.1.2 Virtual Patients

Virtual patients aim to remedy many of the problems with existing simulation techniques. These computerized simulations can represent a wide range of medical conditions and demographics accurately while providing consistent, repeatable experiences for each individual. They can also record data for immediate or post-experience evaluation and feedback. 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) [Hockenberry et al., 2007]. Results indicate that using life-size virtual characters with speech recognition is useful in communication skills education [Planas and Nelson, 2008]. Medical students have used virtual patient scenarios to develop their clinical reasoning skills [Botezatu et al., 2010], communication

skills [Lok, 2006, Stevens et al., 2006, Deladisma et al., 2009], interdisciplinary collaboration [Booth and McMullen-Fix, 2012, Reese et al., 2010] and patient interviewing skills [Johnsen et al., 2007]. Virtual patients have also been used in nursing education to assess clinical reasoning skills [Forsberg et al., 2011] and cultural competence [Sakpal, 2012]. Johnsen et al. [Johnsen et al., 2007] found that the use of a virtual patient to develop medical students' interviewing skills was as effective as the use of actors trained as patients. Similarly, research on the use of virtual patients in medical education found that the most significant role of the virtual patient was in the promotion of the practitioner's clinical reasoning skills [Consorti et al., 2012, Cook and Triola, 2009]. Although virtual patients are capable of simulating many demographics and medical conditions, most virtual patients are adults without any physical abnormalities. We focus on virtual pediatric patients due to their rarity and the difficulty of simulating a pediatric patient through peer interview or standardized patient. The use of virtual pediatric patients was first addressed in [Durham and Alden, 2008], 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 [Durham and Alden, 2008]. 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.

2.2 Assessing Learning in Virtual Patient Scenarios

Virtual patients are increasingly common in nursing education, however, there is significant work to be done in assessing the learning outcomes resulting from their use. Many evaluations center on measuring perceived satisfaction and usability [Harder, 2010], which may indicate that using a virtual patient is a positive experience, but does not indicate anything about how it affected nurse performance or increased knowledge. Other evaluations of virtual patient simulations are based on satisfaction and confidence surveys or knowledge acquisition [Bray et al., 2011]. In their literature review of the use of virtual patients in pharmacy education, Bray et al. [Bray et al., 2011]—established five different categories of assessment: 1) assessment of satisfaction, confidence, and/or self-efficacy, 2) assessment of knowledge, 3) assessment of performance-based skills, 4) assessment of problem-solving abilities, and 5) evaluation of team-based behaviors. An ideal assessment, however, focuses less on knowledge acquisition and more developing the skills necessary for patient care [Bray

et al., 2011].

The importance of developing valid and reliable assessment is a well-accepted tenet of the virtual patient community [Bray et al., 2011, Cook and Triola, 2009]. Certain learning transfer studies show that a rubric-based assessment within a virtual patient system can yield positive learning outcomes [Botezatu et al., 2010]; others have shown that the skills learned in a virtual patient scenario transfer into communication with a standardized patient [Johnsen et al., 2007]. Despite this progress, little research has been done comparing the effectiveness of various virtual patient presentation methods [Cook and Triola, 2009], and few virtual patients offer anything beyond the student-patient interaction to specifically support learning.

2.3 Scaffolding for Learning

Instructional scaffolding is a teaching and learning strategy that provides temporary support to a learner while he or she achieves competency in a task that he or she would not be able to accomplish unaided [Sawyer, 2006]. Traditionally, scaffolding occurs in an apprentice-learner context, with a subject expert aiding a less experienced learner; however, any method of support provided during learning is considered a form of scaffolding [Holton and Clarke, 2006]. Holton and Clarke have classified scaffolding techniques based on agency (who provides the scaffolding) and domain (whether the scaffolding helps teach a concept, or helps teach a heuristic method) [Holton and Clarke, 2006]. There are three possible scaffolding agencies: expert, when a domain expert aids a learner; reciprocal, when peers working together aid each other through their individual strengths; and self, when an individual can apply his or her related knowledge to a new problem. Conceptual scaffolding provides problem-specific aid, while heuristic scaffolding provides principles that are useful for solving generalized problems. Any agency can provide either of the domains. Since the final goal of scaffolding is learner independence, self-scaffolding is especially desirable [Holton and Clarke, 2006].

Educators have successfully applied scaffolding to many domains, including mathematics [Holton and Clarke, 2006], medicine [Lua et al., 2009], and science [Quintana and Fishman, 2006]. Several researchers have applied expert scaffolding to pharmacy students learning to conduct patient interviews. Planas et al. developed a scaffolded communications skills development system that students used as a part of their course on clinical communications [Planas and Nelson, 2008]. This

system provided scaffolding through several means. Students worked in groups, providing reciprocal scaffolding, and received feedback from faculty members and standardized patients, providing expert scaffolding. Students also were required to evaluate their own performance, encouraging self-scaffolding. As the course progressed, support in the form of peer and instructor guidance lessened and students became more competent. Comparing the assessments of a standardized patient interview before the course and one after the course, students performed significantly better ($p < 0.01$) as rated by faculty, peers, standardized patients, and students themselves [Planas and Nelson, 2008]. Stupans et al. developed several tools for scaffolding patient counseling skills, including rubrics for student evaluation, a flexible protocol for counseling sessions, tools to aid in self-critique and peer-critique of interviews, roleplaying, and use of expert scaffolding, especially early in the experiential learning process. Although they have not yet tested its effect on performance, students generally responded positively to the idea of scaffolded opportunities for learning [Stupans and Owen, 2009]. Additionally, several nurse educators have successfully applied scaffolded instruction to bridge the gap between education and clinical practice [Sanders and Welk, 2005, Spouse, 1998, Tilley et al., 2007].

Crossing into the electronic domain, Simmersion has developed several systems designed to teach interaction for specific scenarios, providing aid through a "virtual coach" who gives feedback as users ask questions [SIMmersion, 2013]. Simmersion systems show a videotaped human who acts as the interview partner in a large portion of the screen. The user selects a question from a list of questions to ask the interview partner, then the partner responds through a videotaped clip. A virtual coach provides nonverbal feedback on the question selection through motions such as applauding or shaking her head, and provides text feedback if clicked. One example product is a system that gives people with autism practice in social conversation. With eleven learning objectives, a complex conversational model determines how the interview partner responds to user questions. The system keeps a points-based score for user choices and allows the user to review the interview script. A user study on this system showed that the users enjoyed their interaction with the system but would have appreciated a system with more conversational choices, and that the mean scores improved from the first use of the system to the last use of the system [Trepagnier et al., 2011]. The Simmersion systems provide expert, conceptual scaffolding, since the virtual agent gives choice-by-choice feedback specific to the current conversation. Due to the complexity of the learning objectives and the videotaped interactions, these systems are difficult to extend. Additionally, there

is no leveled scaffoldingthe same feedback is provided to the user by the virtual coach regardless of the user's experience level.

2.3.1 Assessing Communication Skills

To provide learning scaffolding, there must be measurable learning objectives so that the scaffolding system can provide aid and feedback where necessary. For our work, we have chosen to focus on learning objectives related to nursing communication skills due to their importance in nursing practice. While communication skills are vital for any nurse-patient relationship, training a nurse to interact appropriately with parents and children is especially essential. The dynamics related to the nurse-family-child relationship are complicated 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 [Hockenberry et al., 2007]. 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 [Ball et al., 2013]. 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 [Ball et al., 2013].

Although communication skills are difficult to objectively measure, Diers [Diers, 2008] has created a validated rubric to appraise a nursing student's verbal and nonverbal communication along several dimensions. Diers identified several broad components of nursing communication as important: empathy, demeanor, respect, credibility, coherence, and clarity, each of which can be manifested both verbally and nonverbally.

2.4 Authoring Virtual Patient Scenarios

Although simulation training scenarios are useful for student training, they are not utilized to their greatest potential because of the difficulties and time required in developing effective scenarios. There are unique challenges both for nursing educators, who must develop the scenario content, and computer scientists, who must implement the software to make the scenario interactive. The

literature clearly demonstrates a need for improved scenario development for virtual patients.

2.4.1 Nurse Educator Authorship

The role of the nurse educator is primarily to determine the scenario content. In addition to selecting the patient demographic and medical condition, scenario generation requires the careful integration of many elements: 1) the desired learning objectives, 2) the student population and level of expertise, 3) the desired format of the case (e.g. linear, non-linear, or branching), 4) the inclusion of assessment and feedback, and 5) the methods for interactivity [Posel et al., 2009]. Despite nurses' best efforts, Round et al. [Round et al., 2009] noted that difficulties in virtual patient development include inflexibility, non-realistic scenarios, and lack of engaging content in the final products.

2.4.2 Computer Scientist Authorship

Because there are no applications that allow nurses to create fully interactive virtual patients independently, once the content is generated, the burden of implementation falls to computer scientists. The majority of current systems are developed on a case-by-case basis using centralized conversational modeling [Rossen and Lok, 2012]. This process can be time-consuming. Indeed, designing the SIDNIE system took nine months, three months beyond the approximate six months or 200 hours to develop a conversationally accurate virtual patient system [Rossen and Lok, 2012]. Existing commercial products may expedite this process. For example, TheraSim is a commercial system that has been used in the development of virtual patient systems, specifically for teaching clinical skills like treating patients, diagnosing medical conditions and prescribing medications [Rajak and Saxena, 2010, Hadden, 2009]. However, their business model does not allow nursing educators to directly develop or change their own scenarios, still yielding a bottleneck in scenario generation. Unfortunately, the time and expense of these scenarios is seldom justified since once a student works through a scenario, it is no longer as effective for learning. To teach new learning objectives and give a wide range of experiences, there must be a wide range of scenarios.

2.4.3 Existing Authorship Processes & Tools

Centralized conversational modeling is the process behind most virtual patient scenario dialogue generation, in which a 'knowledge engineer' takes the generated scenario from medical

domain experts and other sources and then translates it into simulation data [Rossen and Lok, 2012]. Unfortunately, since the knowledge engineer must process all the obtained input, they inadvertently hinder the dissemination process of this data.

Several researchers have attempted to solve this bottleneck by providing nurses tools to create their own scenarios without the aid of a computer scientist or knowledge engineer. Round et al. [Round et al., 2009] developed an inexpensive method to create multi-path virtual patients; their "virtual patients" were composed of text and photo elements instead of realistic visuals and interaction techniques. Similarly, the Decision Simulation [Benedict, 2010] and OpenLabyrinth [Ellaway, 2010] authorship tools provide great flexibility for nurses, but image-and-text interaction instead of fully interactive patients.

Our review of the computer science domain literature turned up but a single report of an innovative technique: the Virtual People Factory (VPF) approach, which uses an alternative model of "Human-centered Distributed Conversational Modeling" [Rossen and Lok, 2012]. Domain experts (nurse educators) develop a set of questions and answers and input them via a web interface into an virtual patient application generator. They then send the virtual patient application to domain novices (students) so they may question the virtual character [Rossen and Lok, 2012]. When students pose questions the virtual character cannot answer, the error is recorded, which a nurse educator later reviews and adds the appropriate answer. This model produces more complete scenarios in a shorter period of time. Consequently, unlike conventional design methods in which six months and 200 hours are needed to develop a conversational model that is 75% accurate, the Rossen authoring tool can reduce that development time to as little as 15 hours [Rossen et al., 2009].

Although the authoring tool approach greatly reduces the burden of scenario development and requires less attention from a computer scientist, more research and development is necessary. Conversational information cannot be easily reused for other patients, and the model falls short of generating all the components of a realistic simulation. The final product is a static image of a virtual character paired with a "chat" box where a user can type a question and receive a text answer from the virtual patient, or alternatively, a simulation of the virtual patient in Second Life, where the character can display a few generic animations. There are no controls for automated feedback, for drastic changes in either patient appearance or medical condition or adaptation for different learning goals, so any major changes must be implemented by a computer scientist. Additionally,

there is no indication that medical educators were included during the design phase. Though the usability results were positive, the authors admit a possible bias since the participants were research collaborators who had invested substantial time in learning to use the system [Rossen and Lok, 2012]. While scenarios generated with any of these tools may be appropriate for some learning objectives and scenarios, much of the human communication is nonverbal [Thalmann, 2001]; therefore, the communication skills that can be taught with these applications are somewhat limited.

Chapter 3

Initial Virtual Patient Prototype: The Virtual Pediatric Patient System

Our first implementation of a simulated patient is called the Virtual Pediatric Patient System, which is illustrated in Figure 3.1. The Virtual Pediatric Patient System is a simulation of a mother (Mrs. Jones) and her daughter (Sarah) who have come to a medical clinic because Sarah has an earache. The user sees an animated image of the mother and child on a large screen (52") display along with their environment. The user wears a microphone headset and speaks directly to the characters to interact with them. The correct response is then retrieved from a database and executed. Our scenario is based on a written script provided by faculty members in the School of Nursing.

The software implementation can be described as three linked software modules: speech recognition, question matching, and scenario rendering. The overall program is written in C++ with MFC (Microsoft Foundation Class) for supporting interaction between modules. The first module, speech recognition, is implemented using Dragon Naturally Speaking. This module enables a user to create a voice profile for recognition accuracy. We tested two vocabulary options: limited and general. We created the limited vocabulary from words specific to the application. The general vocabulary is provided by Dragon Naturally Speaking containing all the words in the dictionary.

Figure 3.1: A student interacting with the Virtual Pediatric Patient System.

The second module matches speech input to a question in a database. Our corpus of questions and responses are stored in a SQLite database[SQLite, 2012]. We use a variant of the Answers First Algorithm to match the spoken phrase captured by the speech recognition engine to a question stored in the database [Wilson, 2006].

A limitation of the Answers First Algorithm is that a match may not be retrieved if sentences are semantically similar but not syntactically similar. Another problem is that a false match may be retrieved if sentences are syntactically similar but not semantically similar. We can improve matching accuracy by determining and storing many alternative phrasings of a question in the database [Wilson, 2006], but doing this manually is very time consuming. To reduce the amount of time and effort spent generating alternative phrasings of a question, we implemented a sentence generation algorithm to automate the process using the Natural Language Toolkit (NLTK) [Bird, 2006] and Wordnet [Miller, 1995]. First, we took each question from our script and broke it into an array of individual words. Each word was then tagged with its part of speech using the Brill Tagger implemented in NLTK. After that, we removed all stopwords from the sentence, including stopwords defined in the NLTK stopwords list as well as proper names included in our script. Next, we applied the Porter Stemming Algorithm to remove common morphological and inflection endings of words so that synonyms could be found for them in Wordnet-for example, removing the "ing" from "hurting" and mapping the word to its intermediate form "hurt" [Dao and Simpson, 2005]. After the words were stemmed, we used the parts-of-speech tags and the adapted Michael Lesk Algorithm [Banerjee and Pedersen, 2002] to find the correct sense, or definition, of the word in Wordnet. We then used Wordnet to find the lemmas and hypernyms from the synset (the set of all possible related words) provided for that word. Finally, variations of the original sentence were generated by substituting all possible lemmas and hypernyms for each word in the sentence into the same location in the sentence as the original word, and generating every possible combination of those synonyms, yielding a set of semantically similar questions (question set) that is inserted in the database and mapped to the appropriate response.

To match user questions to the correct question set, our question matching algorithm splits the recognized speech into bigrams, which are pairs of words that appear next to each other. In the original Answers First algorithm [Wilson, 2006] the question set with the highest number of matching bigrams was chosen as the correct answer. However, this approach may return an incorrect match when a large question set has a small number of matching bigrams per question, which would yield a

high score although each question matched poorly. To avoid this problem, we determine the matching question set by choosing the question set that has the greatest average number of matching bigrams per question in the question set that contained at least one matching bigram. The corresponding response, stored as a series of actions and sentences to be performed by the characters, is then retrieved from the database.

The final module, scenario rendering, is implemented using DI-Guy [DI-Guy, 2009] for virtual human animation and virtual environment rendering, and Microsoft SAPI SDK [SAPI, 2013] for text-to-speech. Once the responses are retrieved from the database, they are sent to this module so that they may be appropriately spoken and displayed.

3.1 Study Description

After completing our initial prototype, we conducted a usability study where we asked five experienced nursing faculty members to use and evaluate our system. Our goals were to identify problems with the system design, check the stability of the software, evaluate the visual and behavioral fidelity of the simulation, and obtain suggestions of improvement.

3.1.1 Procedures

Each participant began by filling out a consent form and pre-questionnaire. Next, participants completed speech recognition training which consisted of using the short or medium length training module provided by Dragon Naturally Speaking. The participant then sat in a chair in front of the large screen (52") television, which displayed the virtual patients. The participant was then asked to conduct a patient interview with the virtual patients as they normally would. The participants interacted with the virtual patient until they felt they were done with the interview. Each participant then filled out a post-questionnaire and completed a debriefing interview.

3.1.2 Measures

The pre-questionnaire collected data on demographics, occupational status, and computer experience. The post-questionnaire consisted of the System Usability Scale (SUS) [Brooke, 1996], questions about quality of speech interaction, and modified Slater co-presence and presence questionnaires [20]. Through a debriefing interview, we obtained specific information about the [Mortensen

et al., 2002] overall performance of the system. We audio recorded everything the participant said for transcription, and our system logged data related to speech recognition, animation, the underlying conversational model, and every query into the database.

3.2 Results and Discussion

Five experienced nursing faculty members evaluated the system. The nurses were Caucasian females between the ages of 36 and 54. Nurses reported a high level of health care experience as well as a high level of daily computer use (both means above 6, where 7 was frequent experience), but low levels of virtual human experience (mean=2.6 out of 7, sd=1.14, where 7 was frequent use) and daily virtual reality use (mean=1.4 out of 7, sd=0.89, where 7 was frequent use). Participants 1, 3, 4, and 5 interacted with the system once, while participant 2 interacted with the system twice. We gathered questionnaire data for participant 2 once, but we gathered interaction data for participant 2 twice-these instances are referred to as 2a and 2b. Participant 2 interacted with the system using two different settings. Participants 1, 2a, and 2b used short speech recognition training, while participants 3, 4, and 5 used medium length speech recognition training. Participants 1, 2a, 3, and 4 used the system setting for limited vocabulary while participants 2b and 5 used the system setting for full vocabulary, in order to measure whether our limited vocabulary was effective in increasing recognition accuracy.

3.2.1 Verbal Interaction

To evaluate our system's overall performance with respect to the verbal aspect of the interview, we transcribed recordings of the interview so that each line of recognized speech and the system's response was paired with the corresponding line of transcribed speech. We categorized each spoken question-matched answer pair into one of six categories (Table 3.1). Spoken questions that were semantically similar to questions in our database were categorized as either: correct match, which indicates that the virtual patients answered the question correctly; incorrect match but reasonable response, which indicates that the patient gave an incorrect response that made sense in the context of the nurse's question (for example, nodding in response to a yes/no question instead of giving the intended verbal response); or incorrect response, which indicates that the patient gave a response that did not make sense. Questions that were not semantically similar to any question

in the database were classified as: reasonable response; "do not know" response, where the patient responded with a phrase indicating that they did not understand the question; or unreasonable response.

Each nurse asked between 22 and 36 questions (mean=29.5, sd=5.05). 40% of questions were answered reasonably (correct match, reasonable responses, or "do not know" responses). There are several reasons that we believe the reasonable response rate was low. One important observation is that 51% of the questions that the nurses asked were not represented in the database, so the system had no way to respond to over half of the questions asked. Out of the questions asked that we had answers for in the database, only 16% of them were answered correctly. This could have resulted from lack of accuracy in speech recognition and insufficiency of our conversation model.

3.2.2 Speech Recognition and Conversational Model

By comparing transcribed speech to recognized speech, we found that overall speech recognition correctly recognized 66% of the words nurses used. Nurses with the setting for limited vocabulary had recognition accuracy of 60%, while the recognition accuracy for nurses with the setting for a full vocabulary was 86%. Training length had a negligible effect on recognition accuracy (66.17% for short training and 67.39% for medium training).

We originally chose a limited vocabulary in hopes of improving accuracy, assuming that most words that a nurse used would be included in the automatically generated sentences. Our analysis shows this is true-82% of the words that nurses used were in our limited vocabulary. However, looking at the spoken lines in comparison to the recognized lines, the words that nurses used that were not present in the vocabulary were mapped into incorrect words. These mismatches caused the recognition rate to be much lower for participants in the limited vocabulary than in the full vocabulary. A converse problem was observed using the full vocabulary: words were often mapped to homonyms that made no sense in context of our application. We feel that the benefits of a higher recognition rate through a full vocabulary outweigh the disadvantages, and in future revisions we plan on using a full vocabulary. Out of the 82% of words that nurses used that were in our database, 72% of the words were in our original script, while the additional 10% were added through our sentence generation algorithm. In order to make a manageable and accurate word set we used a subset of the synsets, but in future iterations we may gain better vocabulary coverage by using a larger subset. In addition to better question matching in the database, this larger vocabulary could

Table 3.1: Categorized questions-answer pairs by participant. Reported values are percentages of the total questions the participant asked.

Participant	Question in Database			Question not in Database		
	Correct	Reasonable	Incorrect	Reasonable	Don't Know	Unreasonable
1	14	6	26	18	6	29
2a	19	3	25	0	11	42
2b	11	7	39	14	14	14
3	18	0	27	9	14	32
4	17	0	30	3	7	43
5	19	11	11	11	19	30
Mean (sd)	16 (3.20)	5 (4.32)	26 (9.07)	9 (6.74)	10 (4.88)	32 (10.56)

make limited vocabulary matching in speech recognition more accurate.

Since our question matching algorithm is bigram based, we also evaluated speech recognition and sentence generation in terms of bigram matching. Speech recognition correctly recognized both consecutive words in a spoken bigram 53% of the time. Training length only had a small effect on bigram recognition accuracy (51.58% for short training and 54.92% for medium training). 31% of the bigrams that speech recognition recorded (whether correctly recognized or not) were present in our database. These percentages are a sharp drop from the word-based matching statistics. We chose a bigram-based matching model to help provide context for our word matches. However, this observation suggests that in the next iteration of this software we should consider matching our questions using word-by-word matches instead of bigram matches. Additionally, we noticed that many questions asked with a correct keyword were not answered correctly due to bigram matching. For example, a nurse asked "Is Sarah allergic to any medicine?", but because the bigrams "Sarah allergic" and "allergic to" were not in the database, the question did not find a match, although "allergic" would have been a clear keyword choice for the matching sentence in our database. This is also a limitation caused by our sentence generation algorithm. Since our current algorithm only replaces synonyms of words in the same position in the sentence of the original word, all of our question variants are syntactically identical which makes the bigrams syntactically similar as well.

3.2.3 Nurse Feedback

All nurses expressed frustration with the verbal interaction and commented that they felt that they could not interact naturally with the patient because they had to rephrase questions unnaturally. This suggests that we may need to provide some guidance or prompting as to what

questions the virtual patient is able to answer. Most nurses also commented that they felt that the system was unresponsive. In most cases the database response time was less than one second, but because many of the character's responses were nonverbal signs leading up to a verbal answer, it seems that nurses did not realize that the system was responding to their question. Many nurses asked questions in succession without giving the system enough time to process their input and respond. One limitation of our question matching algorithm is that the length of the recognized speech string affects the processing time for finding an answer, so as nurses tried to accommodate for the system's unresponsiveness by stringing several questions together, the processing time increased.

Despite speech recognition performance, all of the nurses still expressed that they would prefer to interact with the system through speech because that is the way a real patient interview works. On a scale from 1 (not at all) to 7 (very much), all nurses ranked the usefulness of talking to the system as 4 or above (mean=5.6, sd=1.14).

3.3 Lessons Learned

An effective virtual patient should provide an efficient pathway for students to be able to achieve some level of competence in interviewing techniques before they reach the clinical setting. Our first approach, the Virtual Pediatric Patient System, did not provide that pathway. Our usability study showed that even experienced nurses had difficulty using the system, since there was no guidance for which questions our system could answer, and because of problems with visual and behavioral fidelity. Additionally, to give a user feedback on his or her performance required the attention of a nurse, since there was no automated scoring system. To better support learning and reduce difficulty, we decided to build a scaffolded system for interview practice where a student nurse selects questions from a question bank to interact with the virtual patient. Although speech interaction was deemed important by the initial user group, interaction cues and proper feedback are more essential to ensure positive learning outcomes.

Chapter 4

SIDNIE: Scaffolded Interviews Developed by Nurses In Education

To improve upon the virtual patient system, 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. Extending the success of the apprenticeship model already present in nursing education, we designed this system to provide leveled heuristic scaffolding through expert agency while supporting maximum extensibility. Our virtual nurse serves as the expert scaffolder, teaching heuristic knowledge by providing information on criteria that pediatric patient interviews should meet and broad categories that an interview should cover. The virtual nurse provides scaffolding by giving instruction on the criteria and categories, and by giving feedback on user selections. We encourage progressive learner independence by providing several levels of scaffolding, ranging from heavy guidance to minimal feedback. Additionally, we preserve extensibility through our database-driven design, criteria-based scoring, and virtual patients, whose actions and speech can be easily modified.

SIDNIE 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 guides the user through several scaffolded practice opportunities and provides feedback on user choices. The user conducts an interview with the virtual patients by selecting questions from a preset

list of questions developed by experienced nurses, and the virtual patients respond appropriately. The user can also view interview playbacks and score them as if they were an instructor, encouraging knowledge application to new situations. SIDNIE runs on a standard desktop computer and a single monitor, and the user interacts with the system using the mouse.

4.1 Scenario Development

We worked with experienced nursing faculty to develop a patient-nurse interaction scenario for a five year old child with an earache. The nurses generated a set of possible questions and scored each question according to two criteria important for pediatric interviews: age appropriate, and unbiased. Age appropriate in this context primarily means that questions addressed to the child used language children would understand, and that necessarily complex questions were addressed to the parent. For example, "Hello Sarah, I am the RN who will be your nurse" is not age appropriate, since a child would not know that RN stands for "Registered Nurse." Unbiased questions do not make implicit assumptions or lead the patient to respond in a particular way. For example, "Sarah, your tummy hasn't been hurting, has it?" leads the child to respond negatively, and is therefore biased. Additionally, the nurses categorized the questions into broad categories and subcategories, and provided us with the subset of categories that are essential for a nurse to cover in every patient interview. The subset of categories and subcategories that are essential are: Introduction/Introducing Yourself, Introduction/Confirming Patient Identity, Chief Complaint, History of Present Illness, Related Systems, and Past Medical History. This detailed structure enables us to provide feedback to users on their question selections and interview thoroughness. This meets our goal of providing heuristic scaffolding, since these criteria and categories can be applied to any pediatric patient interview.

4.2 Sidnie

A male nurse named Sidnie serves as our virtual nurse educator during the interview simulation. Sidnie's role is to provide instruction on how to use the other interaction modules, to give feedback to the user on his or her performance, and to demonstrate patient interviews. Sidnie is represented by a static image of a male nurse in the lower right hand corner of the screen. We chose

not to animate Sidnie or support additional interaction with him because we felt that an additional animated character would distract from the patients' behaviors and animations. Sidnie has a quote bubble above his head that is used to display instructions or give feedback.

4.3 Tutorial

When each session begins, Sidnie introduces himself through text and guides the user through a tutorial that outlines how to use the level using screenshots with accompanying text. The user must click "Next" to proceed through each slide until the tutorial is finished. Sidnie can give additional text or image instructions according to what is present in the database.

4.4 Electronic Health Record

After the tutorial, our system displays a simple electronic health record (EHR) for the patient. The EHR features information on the patient's current statistics including a photo of the patient, the patient's birthday, weight, temperature, immunization record, medications, allergies, parents' names and contact information, and reason for the visit. The user can access this information using buttons to select categories of information. Once the user has reviewed the medical record, he or she clicks a button to proceed with the scenario.

4.5 Pediatric Patient Interview

During the scenario, the majority of the screen displays the virtual patients in their environment. For our current simulation, a mother and her five year old daughter sit in a pediatric patient room. In an idle state, the virtual characters sit quietly. When responding to questions, the patients perform appropriate speech and animations. Depending on the chosen scaffolding level, the user may either conduct a patient interview with guidance or may observe an interview and score it for the chosen criteria.

Figure 4.1: 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.

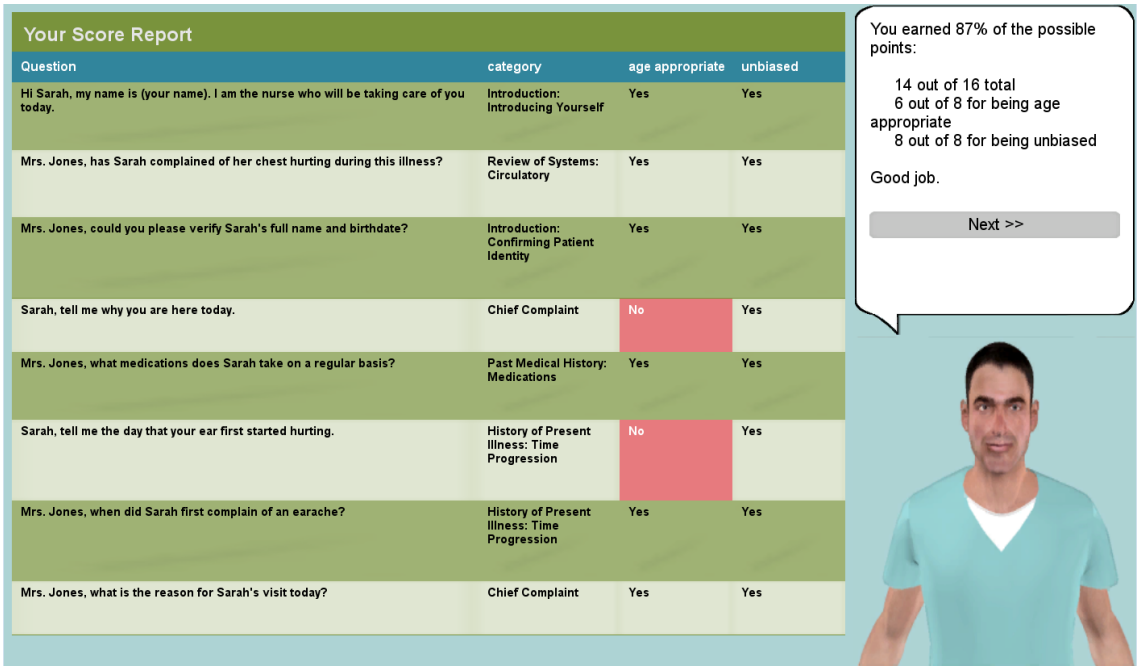


Figure 4.2: When summative feedback is provided, at the end of the scenario, a chart is displayed with the score for each question the user asked.

4.5.1 Conducting an Interview

Our database stores a set of questions that the user may ask, and the corresponding character responses. For ease of use, we display the questions according to their categories and subcategories in the question selection interface. The question selection interface is displayed at the bottom of the screen when the user is permitted to conduct an interview with patients. On the left side of the screen, colorful tabs permit the user to select top level categories, which are dynamically loaded in from the database. Users may select subcategories using a set of tabs along the top of the list of questions. Under the tabs, we display the list of questions belonging to the selected category and subcategory. We randomize the order the questions are displayed with each run of the application so that repetitive use is still slightly varied. To select a question, the user simply clicks on the displayed question. The question is highlighted, then the characters respond appropriately with speech and/or animation. After the characters respond, the user may proceed to select questions until the session is complete. See Figure 4.1 for a screenshot of the interview interface.

4.5.2 Grading an Interview

Sidnie may also prompt the user to grade another interviewer's questions to encourage the user to apply his or her knowledge to a new problem. In this module, the application retrieves a prerecorded sequence of questions from the database. Sidnie "reads" a stored question using text to speech, then the virtual patient and/or parent responds with the corresponding answer stored in the database. After the characters finish responding, in his quote bubble, Sidnie prompts the user to grade the question on the stored criteria, and to categorize the question into the correct category & subcategory. We provide a checkbox for each criteria, where the user should check the checkbox if it meets the criteria, and then a scroll down list, where the user must select a category before proceeding. After submitting the score, the interview playback proceeds.

4.6 Feedback Mechanisms

To promote user learning, we provide two types of feedback for the user depending on the selected scaffolding level. Immediate feedback provides selection-by-selection scoring so that the user can correct his or her mistakes immediately during the interview, and is used in the lower scaffolding levels. Summative feedback is provided at the end of each scaffolding level and provides the user with an overall score and overview of his or her performance for the entire scenario.

4.6.1 Immediate Feedback

When the user is conducting an interview, Sidnie may give immediate feedback on user question selection in his quote bubble. When the user selects a question, we query the database to see whether the question meets the stored criteria. We use the query results to generate a string that Sidnie may immediately display in his quote bubble to indicate how the user performed. For example, for a question the user selected that meets both criteria, Sidnie may say, "Good job! The question you asked was unbiased and age-appropriate." When the user is scoring an interview, Sidnie also may provide immediate feedback on the submitted score. When the user submits a score, we query the database to find the actual question's scoring and category, then compare them to the user selected characteristics and build a string to provide feedback. For example, the correct category was chosen but both criteria were scored incorrectly, Sidnie may say, "You selected the appropriate category for this question: 'Introduction: Confirming Patient Identity.' You scored the

question incorrectly as: unbiased when it was not unbiased, not age appropriate when it was age appropriate."

4.6.2 Summative Feedback

At the end of each session, the user may receive summative feedback in chart form as well as an overall percentile score for his or her performance. Each row in the chart represents a question the user asked, along with its scoring data. If the user conducted the patient interview, we highlight the cells for the scored criteria based on whether the selected questions met the criteria. Pink cells show the user that the question did not meet the criteria, while green cells show that the question met the criteria. Additionally, we display a pink row with the category name for each essential category that the user neglected. If the user scored a patient interview, we highlight the cells based on whether they scored the question correctly. Pink cells show the user that they made a mistake in scoring or categorizing the question. In the summative feedback screen, Sidnie displays percentage and proportion scores in his quote bubble. If the user conducted an interview, we calculate the percentage score and proportion score by counting how many questions met the criteria and how many of the essential categories the user covered. We generate speech for Sidnie based on those scores. For example, for a user who asked seven questions, and covered each essential category, but did not meet the scoring criteria for every question, Sidnie may say, "You scored 12 out of 20 points, for a score of 60%. You scored 6 out of 6 points for category coverage, 5 out of 7 points for being unbiased, and 1 out of 7 points for being age appropriate." We calculate scores similarly for users who score an interview, except we score them for choosing the correct category and for correctly grading the question for the criteria instead of choosing questions that met criteria and covered essential categories. Figure 4.2 illustrates an example summative feedback screen.

4.7 Scaffolding Levels

Our goal in this system is to provide leveled scaffolding so that an inexperienced student may progressively gain enough confidence, knowledge, and simulated experience to ease the transition into actual patient interviews. Currently we have implemented four scaffolding levels for our simulation designed to progressively remove scaffolding until the student can perform an interview without aid.

We based our scaffolding levels on those used for in-person pharmacology interviews [Planas

and Nelson, 2008, Stupans and Owen, 2009]; however, our approach is innovative in that we use one set of criteria to score each question, so that students learn overall concepts to be applied for each question in addition to the steps needed for a complete interview.

4.7.1 Level 1: Baseline Measurement

To enable measurement of user progress, we first require the user to conduct a patient interview using our system without any training or guidance. During the tutorial phase, Sidnie explains how to select a question using our system. The user reviews the medical record and proceeds with the interview, selecting as many questions as he or she desires. Sidnie does not give immediate feedback on the question choices. When the user is finished with the interview, the session concludes without any summative feedback.

4.7.2 Level 2: Guided Interview with Immediate Feedback

The second level begins the scaffolded learning by guiding the user through a short pediatric patient interview. During the tutorial phase, Sidnie explains how to select a question using our system, and then describes the criteria that the questions should meet and the categories that an interview should cover. The user initially reviews the medical record then continues to the guided interview. The system preselects the first essential question category in the question selection interface and disables the user from changing the category. Then Sidnie instructs the user to select a question. When the user selects a question, the patients respond, then Sidnie gives immediate feedback on whether the question met the criteria. If the question did not meet each of the criteria, Sidnie informs the user of which criteria were not met and instructs the user to try again. The user must continue to select questions until he or she chooses a question that meets both criteria. Then, the system advances to the next essential category. The process continues until each essential category is covered, then shows summative feedback to the user.

4.7.3 Level 3: Grading an Interview

The third level requires the user to score a prerecorded interview by watching and listening to each question and response, then identifying the category. The tutorial phase explains the interview playback process and the grader interface. The user then views Sarah's electronic medical record

and proceeds to the interview. As described in the previous section on grading an interview, the user grades the interview question by question, and Sidnie gives immediate feedback on his or her grading, but does not require the user to correctly grade the question before moving to the next question. Once the user finishes scoring the interview, he or she receives summative feedback on the scoring.

4.7.4 Level 4: Evaluation

Finally, the fourth level evaluates the user's learning by providing unguided interaction similar to Level 1. The tutorial phase again explains the system and the scoring criteria, and then the user proceeds through the interview by selecting as many questions in as many categories as they wish. The user clicks a button when he or she is finished. Lastly, they receive summative feedback on their interview based on their category coverage and whether their questions met the criteria.

4.8 Implementation Details

Our system is implemented primarily in Unity 3D [Unity3D, 2012] as a set of related modules implemented in Boo or C#. We created the virtual environment using Blender [Blender, 2012] and imported it into Unity. We rigged and animated the characters using Poser Pro [Poser, 2012] and [Blender, 2012]. Using Unity's plugin interface, we built a C++ plugin as an interface to SAPI 5.3 [SAPI, 2013]. We used IVONA voices [IVONA, 2012] for our virtual characters: Sarah's voice is Ivy, Mrs. Jones' voice is Kendra, and Sidnie's voice is Joey. In addition to the animations triggered by responding to user questions, our characters exhibit baseline animations such as breathing and blinking.

Although our virtual characters and environment are static at this point, the remainder of our program is database-driven, providing flexibility to change scenarios, scoring criteria, instructions, and required categories. We store our scenario data in a SQLite 3 database [SQLite, 2012] and access it through Unity's plugin interface. Additionally, each question selection and scoring action is timestamped and recorded in the database, keyed on the user ID and the scaffolding level for easy data processing.

4.8.1 Character Response Queueing

When a user clicks on a question, we retrieve the answer the characters should give from the database. We store the answer as a sequence of animations and text to speech, with each record in the response representing one animation or line of speech. Each action is accompanied by an identifier that associates it with one of the virtual characters, and a delay that denotes when in time the action should occur. Within our Unity application, we keep a queue of these actions with their delays. While the queue still contains items, we use a coroutine to keep track of elapsed time and check whether it is time to perform an action. When it is time to perform an action, we call the appropriate function and remove the action from the queue.

Animating a character is a simple Unity function call to trigger a preset animation. When a character's response requires speech, the process is more complex. Because timings in the database may be inexact, and creating many SAPI instances causes performance problems, we also keep a text-to-speech queue that holds any text to speech to be rendered. When a new question is selected, we count the number of speech items contained in the response, then send that number ahead to the plugin so that it can indicate to Unity that it is busy until there are no more items to be spoken. Using a coroutine, while there is speech in the queue, our application checks to determine whether the text-to-speech engine is currently rendering speech. When the engine is not busy, it sends the next item in the queue, then waits for it to be non-busy again. To keep the count of items to speak up to date, within the plugin, we handle the SAPI event for finished speech and use it to decrement the counter.

We send the string to SAPI via the C++ plugin, along with parameters to select an appropriate voice based on the character's age and gender. This increases our system's extensibility since it would cause the most appropriate voices to be selected on a system that does not have our chosen voices installed. Since Unity's default behavior is to call a plugin as a part of its own thread, and SAPI voice loading events are blocking function calls, within our plugin we launch a thread to render the text to speech so that the Unity application may continue to render the virtual characters with appropriate animations and lipsyncing. Within the launched thread, the plugin loads the appropriate SAPI voice and renders the string through text-to-speech. Inside the plugin we store the current viseme for each character (mouth shape) by handling SAPI viseme events as the text to speech is rendered. In the Unity application's update function, we query the plugin to gather each

character's current mouth shape. If the mouth shape has changed since the previous update, we trigger an animation to change the character's mouth shape to match the speech SAPI is rendering, using a modification of the Metamorph script [Rezzable, 2012] to apply shapekey morphs to our characters' faces.

In order to encourage the user to be attentive to the characters' responses, we disable the question selection interface while the character is performing animation or speaking. We do this by checking to determine whether the response queue is empty and whether the text to speech plugin has more speech to render. While the queue is nonempty or text to speech has more lines to render, we overlay the question selection interface with a translucent screen and place text over the interface and in Sidnie's text bubble to indicate that the user should wait for the characters to finish their response. After the character has finished her response, the question selection interface is enabled and Sidnie provides appropriate instructions for the scaffolding level.

Chapter 5

Evaluation and Refinement of SIDNIE

Once we implemented an initial proof-of-concept for the SIDNIE system, we evaluated several aspects of SIDNIE in regards to student performance. We then used student suggestions and observations of their interaction with the system to make modifications for better usability and improved learning outcomes.

5.1 The Effects of Scaffolding on Learning Outcomes

We conducted an experiment to gather initial impressions on SIDNIE's usability and to measure learning outcomes for a student completing all three scaffolding levels. Each of the scaffolding levels used the same scenario: a mother with her five year old daughter, who has an earache.

Participants completed a demographics questionnaire and read the informed consent form the morning before the experiment. When the participants came to the experiment site, we greeted them then gathered their demographic questionnaires, which included questions about previous computer usage, experience in healthcare, and exposure to virtual environments. The participant then signed the informed consent. We seated the participant at a desktop computer and began the SIDNIE system on scaffolding level 1. We asked the participant to view the beginning tutorial screens, which instructed them on how to use the system but included no information on the required

Table 5.1: Demographic information for the study participants (1= no experience, 7= very experienced)

	Mean	SD
Healthcare experience	3.80	1.37
Experience with children	4.40	1.96
Experience in a healthcare setting	3.60	1.77
Experience with children in healthcare	2.80	1.78
Computer use	6.47	0.74
Virtual human exposure	2.87	1.51
Virtual reality exposure	2.20	1.26

question categories or the scoring criteria, then to conduct an interview with the patients to the best of their ability, and to let us know when they had finished. After the participant completed the initial interview, we began the SIDNIE system on level 2. We told the participants that SIDNIE would guide them through their interview this time, and that they would receive feedback on their choices. The participants then proceeded through the guided interview and notified the experimenter upon completion. The participant continued through levels 2 and 3 similarly, with the experimenter briefly commenting on each level then starting the application.

After completing every level, the participant filled out the System Usability Scale (SUS) [Brooke, 1996] and was given the opportunity to provide any written feedback. Finally, the participant completed an interview with the experimenter, where we asked questions about his or her opinion of the system.

Our primary hypothesis was that participant performance in terms of essential category coverage and choosing questions that met the criteria of age appropriateness and unbiasedness would increase between the level one baseline measurement and the level four evaluation measurement. We also expected participant performance to increase between each scaffolding level, and expected participants to report our system to be usable.

Fifteen undergraduate students seeking their Bachelor of Science in Nursing (BSN) at Clemson University participated in our experiment as an optional extra credit assignment for their Nursing Care of Children class. All of the participants already have received a bachelors degree in another field. Fourteen of the fifteen students are a part of the accelerated second degree program, while the remaining participant is a traditional BSN student.

Twelve of the participants were female, and three of the participants were male. Participants' ages ranged from 23 to 48 (mean=30.13, sd=7.28). Participants reported a high level of

computer use. Participants reported some level of healthcare experience, experience with children, and experience in a healthcare setting, and somewhat lower levels of experience with children in healthcare, virtual human exposure, and virtual reality exposure. See Table 5.1 for detailed demographic information. When asked about their previous experience with virtual reality, virtual humans, or avatars, participants cited their experience in the nursing simulation lab, which has physical dummies that exhibit instructor-controlled symptoms. Most of the students also participated in an experiment in the previous semester that required them to take vital signs for a virtual human in a virtual hospital displayed on a desktop computer and record the results in an electronic health record.

Technical difficulties caused five sessions to terminate prematurely. We gave these participants the option to try the session again, or to continue on with the next level. Three participants chose to redo the session, while two participants continued onto the next level. We were able to recover sufficient data for all but one participant, who chose not to complete level three. We excluded her data from the quantitative analysis but kept her usability data. We recovered the three participants' data who made a second attempt at level completion by splicing together their data from the two session attempts. For the final participant who chose not to retry a level, we considered her data in terms of proportions only, to give a fair basis for comparison to others who completed the level.

Our primary interest in this evaluation was to determine whether user scores improved between the level one baseline measurement and the level four evaluation measurement. Levels one and four were identical except for the instructions during the tutorial phase and the summative feedback at the end of the interview in level four. We calculated the percentage of questions the user asked that were age appropriate, the percentage of questions the user asked that were unbiased, and the percentage of the essential categories that the user covered. Using a two-tailed t-test for paired samples, we found significant differences in each of the scores for the criteria, indicating that participants' performance improved as a result of the use of our system ($p < 0.01$ for each criteria). We did not find a significant difference in the percentage of categories covered. See Table 5.2 for detailed statistics.

Table 5.2: Percentage means and standard deviation for each scaffolding level.

	Age appropriate		Unbiased		Category Coverage	
	Mean	SD	Mean	SD	Mean	SD
Level 1	74.12	12.04	70.65	10.76	91.67	19.34
Level 2	94.22	10.14	93.46	11.18	n/a	
Level 3	82.14	8.75	80.10	13.19	n/a	
Level 4	93.52	9.72	90.26	8.39	100.0	0.0

5.1.1 Progress through Scaffolding

To gauge which of our scaffolded levels were effective for learning, we analyzed whether participants' scores improved over each level. Levels two and three are significantly different from levels one and four, so it is somewhat difficult to make a direct comparison in user performance. We chose to compare the levels based on the two criteria: age appropriateness, and bias. Comparing category coverage was not possible since level two forced all essential categories to be covered, and level three played back a predetermined interview that covered all essential categories. We calculated the criteria percentages for levels one, two, and four for each of these criteria by finding the proportion of questions the user asked that met the criteria. For level three, since users were scoring an interview instead of conducting an interview, we calculated the proportion of questions correctly scored in regards to that criteria.

Using a two-tailed t-test for paired samples, we found significant differences in scores between each of the consecutive scaffolding levels (every $p < 0.01$), however, the mean scores were not always increasing, as we would have expected. The scores for level three were significantly lower than the scores for level two and level four (see Table 5.2). For nonconsecutive levels, there was no significant difference between levels two and four, but a significant difference in both criteria between levels one and three (both p's < 0.03), with higher scores in level three.

5.1.2 Usability and Qualitative Feedback

Overall, participants seemed to find the system easy to use. Our system received an average SUS score of 80.33% (sd=12.20%). The highest scoring prompts were, "I didn't need to learn a lot of things before I could get going with this system" and "I didn't think there was too much inconsistency in the system", both with means of 5.6 out of 6 and the lowest standard deviation of any question as well (sd =0.63). The lowest scoring prompt was "I felt very confident using the

Table 5.3: Percentage means and standard deviation for each scaffolding level.

	Mean	SD
I think that I would like to use this system frequently	4.2	0.94
The system was not unnecessarily complex	5.13	1.25
I thought the system was easy to use	4.13	1.85
I don't think I would need the support of a technical person to be able to use this system	5.13	0.92
I found the various functions in the system were well integrated	4.4	1.18
I didn't think there was too much inconsistency in the system	5.6	0.63
I would imagine that most people would learn to use this system very quickly	4.67	1.59
I didn't find the system very cumbersome to use	5.4	0.83
I felt very confident using the system	3.93	1.75
I didn't need to learn a lot of things before I could get going with this system	5.6	0.63

system" with a mean score of 3.93 out of 6 (sd=1.75). See Table 5.3 for other detailed results. When asked, every participant said that they thought they and their peers would use our system if it were made available to them. Comparing this technique of interview practice to other techniques (such as roleplay, standardized patients, or reading case studies), all participants felt they would have performed the same or worse using another practice technique. Reasons cited for improved performance using our system included less performance anxiety, being able to select from provided choices instead of coming up with questions on their own, being able to get immediate feedback, and being able to take our system more seriously than roleplay with their peers.

Three of the fifteen participants said that the child's symptoms were not realistic for a five year old with an earache. Two participants were not sure since they did not have much experience with children. The remaining ten participants said the child's symptoms were realistic, citing examples such as the data in the electronic medical record, the child pulling her ear, not responding to questions that were not age appropriate, and responding nonverbally. Participants suggested that the patients should have appropriate facial expressions, and that Sarah could be more whiny, fidgety, or irritable. Three participants thought that the mother's voice sounded "mean", but one participant commented that a disagreeable parent is realistic in a clinical setting.

When asked what they liked about the system, nine of fifteen participants mentioned some aspect of the program that related to the scaffolding, including guidance and feedback, time to think about choices, and being able to try multiple questions in an interview with no risk of offending a patient. The most common aspects that participants disliked were the mother's voice and the scoring

interface.

5.1.3 Discussion

Our primary hypothesis was confirmed–participants' performance did improve from the baseline to the evaluation. However, the scores did not monotonically increase over each scaffolding level as we expected, and determining which scaffolding levels are effective for learning is difficult. There was a significant difference in performance between levels one and two, where the only training received was the tutorial preceding level two, and the immediate feedback for each question selected in level two. At level two, it seems that the scores hit a "ceiling", with participants scoring upwards of 93% for each criteria. In fact, nine out of the fifteen participants showed perfect scores in the second level. These scores do not leave much room for improvement in the similar task for level four, and in fact, we did not find significant differences in the performance between level two and level four.

The "ceiling effect" in the scores for levels one, two, and four may be explained by considering literature in the learning domain such as the Revised Bloom's Taxonomy [Forehand, 2010, Anderson et al., 2005]. Our goal is to encourage the highest levels of learning, which in the new taxonomy are titled "creating" and "evaluating". In the first and fourth scaffolding levels, we aimed to permit users to "create" an interview scenario with limited choices. However, because there were only a few choices in each category and subcategory, and a participant was potentially exposed to each of those questions and their corresponding responses and scores multiple times, by the time the user reached the fourth level, the task became a low-level learning exercise of "remembering" or "understanding", or at best, "applying" the meaning of the criteria and categories to select appropriate questions. In retrospect, given our limited question base, a more appropriate baseline and evaluation may have been to require the user to score an interview before and after some training. The scoring task reaches the "evaluating" level of the learning taxonomy, and the lower percentage scores in the third level of scaffolding can possibly be attributed to the higher level of learning required.

The lower scores for the third level are also likely the result of usability problems. Several participants were confused by the scoring interface. Since "unbiased" is a negative word and was paired with a checkbox, several participants asked questions like, "Do I check this box if it is biased, or unbiased?" In future iterations of the design, we will change to a radio button interface where the user can select "Yes" or "No" to clarify the scoring. Additionally, several users mentioned that

they did not realize until far into the session that they could scroll down to select categories besides those in immediate view. Unity3D's default scrollbars are dissimilar to standard scroll controls and do not feature up and down arrows, so we will modify their appearance to match more traditional interfaces, or we will change the scrollable view to a drop-down box. Lastly, although our system currently supports a question belonging to only one category, several students felt that some of the questions could have fit in multiple categories. In the future, we will consider allowing questions to be part of multiple categories.

Although there is always room for improvement, our SUS scores as well as the feedback from our debriefing interview shows that our system has good usability, confirming our final hypothesis. The lowest scoring SUS question was about user confidence. Although we did not gather data on specific sources of anxiety, since each participant reported that they and their peers would use the system if it were offered to them, and that they would expect similar or worse performance using another interview technique, the low score for this question may reflect a user's lack of confidence in their ability to conduct pediatric interviews instead of their confidence in actually using the system.

5.2 The Effects of Visual and Interaction Fidelity on Learning

In addition to evaluating the learning outcomes of SIDNIE due to its scaffolded approach, we were interested in determining the visual and interaction fidelity requirements necessary for learning. We conducted two experiments, one investigating visual fidelity, and the other investigating interaction fidelity. 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

Figure 5.1: Student interacting with SIDNIE via speech recognition.

knowledge. An example of a student talking to the system is shown in Figure 5.1.

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.

5.2.1 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" [Forehand, 2010]. 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) [Brooke, 1996] 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 [Mortensen et al., 2002]. 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.

5.2.2 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 were part of an accelerated BSN program for individuals who already had a bachelors' degree. Participants reported a middling level of health care experience (mean=3.90, where 1=none and of 7=a great deal, sd=1.66), a high amount of daily computer use (mean=6.05, where 1=none and of 7=a great deal, sd=1.12), and a low amount of virtual reality exposure (mean=2.24, where 1=none and of 7=a great deal, sd=1.37). Two participants were male, while 19 participants were female. There were no significant demographic differences between participants in the two conditions.

For the visual fidelity study, 54 freshman nursing students participated (27 in each condition), ranging from 18 to 20 years old (mean=18.24, sd=0.47). Participants reported a low amount of health care experience (mean=2.15, where 1=none at all and 7=a great deal, sd=1.00), a high amount of daily computer use (mean=5.91, where 1=none and of 7=a great deal, sd=1.09), and middling exposure to virtual reality (mean=2.55, where 1=none and of 7=a great deal, sd=1.18). Two participants were male while 52 participants were female. There were no significant demographic differences between participants in the two conditions.

Table 5.4: 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. There were no interaction effects in this test.

				Pre-test		**Post-test**		**Condition Effects**		**Overall SIDNIE Effect**	
	Criteria	**Fidelity**	**N**	**Mean**	**SD**	**Mean**	**SD**	**F-test**	**p**	**F-test**	**p**
Interaction	**Unbiased**	**High**	8	97.50	3.45	100.00	0.00	F(1,16)=0.62	0.44	F(1,16)=6.77	<**0.01**
		Low	10	98.67	2.81	100.00	0.00				
	Age appropriate	**High**	9	81.67	20.68	95.31	9.33	F(1,17)=0.55	0.47	F(1,17)=13.47	<**0.01**
		Low	10	74.00	21.42	94.50	6.94				
Visual	**Unbiased**	**High**	25	99.73	1.33	100.00	0.00	F(1,49)=3.10	0.08	F(1,49)<0.01	0.99
		Low	26	99.42	2.05	99.17	2.37				
	Age appropriate	**High**	25	68.89	18.84	94.31	13.40	F(1,48)=1.23	0.27	F(1,48)=81.23	<**0.01**
		Low	25	65.73	15.02	90.24	12.30				

Table 5.5: 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.

		Pre-test		**Post-test**		**Experiment Effect**	
Experiment	**Criteria**	**Mean**	**SD**	**Mean**	**SD**	**F(1,67)**	**p**
Interaction	**Unbiased**	98.15	3.07	100.00	0.00	2.04	0.16
	Age appropriate	77.63	20.86	94.88	7.93	4.20	<**0.01**
Visual	**Unbiased**	99.58	1.73	99.58	1.73		
	Age appropriate	67.31	16.93	92.28	12.89		
						SIDNIE Effect	
						F(1,67)	**p**
Overall	**Unbiased**	99.20	2.23	99.69	1.49	6.41	<**0.05**
	Age appropriate	70.15	18.53	92.99	11.74	64.18	<**0.01**

5.2.3 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.

5.2.3.1 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 under the high fidelity condition also scored 100% 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 5.4 and 5.5.

5.2.3.2 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 5.6 and 5.7.

Table 5.6: 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. There were no interacton effects in these tests.

				Pre-test		**Post-test**		**Condition Effects**		**Overall SIDNIE Effect**	
	Criteria	**Fidelity**	**N**	**Mean**	**SD**	**Mean**	**SD**	**F-test**	**p**	**F-test**	**p**
Interaction	**Unbiased**	**High**	7	88.57	15.74	88.57	15.74	F(1,13)=0.60	0.45	F(1,13)=3.42	0.09
		Low	8	92.50	10.35	75.00	17.73			F(1,13)=3.20	0.07
	Age appropriate	**High**	7	80.00	23.09	88.57	15.74	F(1,13)=0.30	0.59		
		Low	8	72.50	23.75	87.50	10.35				
Visual	**Unbiased**	**High**	27	86.67	15.69	94.07	14.48	F(1,50)=1.65	0.20	F(1,50)=2.56	
		Low	25	84.80	17.59	87.20	18.15			F(1,50)=11.66	**<0.01**
	Age appropriate	**High**	27	73.33	24.18	86.67	20.00	F(1,50)=0.03	0.86		
		Low	25	72.00	24.49	86.40	19.77				

Table 5.7: 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.

		Pre-test		**Post-test**		**Experiment Effect**	
Experiment	**Criteria**	**Mean**	**SD**	**Mean**	**SD**	**F(1,65)**	**p**
Interaction	**Unbiased**	90.67	12.80	81.33	17.67	0.40	0.53
	Age appropriate	76.00	22.93	88.00	12.65	0.24	0.62
Visual	**Unbiased**	85.77	16.49	90.77	16.55		
	Age appropriate	72.69	24.10	86.54	19.69		
						SIDNIE Effect	
						F(1,65)	**p**
Overall	**Unbiased**	86.87	15.78	88.66	17.14	0.47	0.5
	Age appropriate	73.43	23.71	86.87	18.27	10.14	**<0.01**

5.2.3.3 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 5.2.

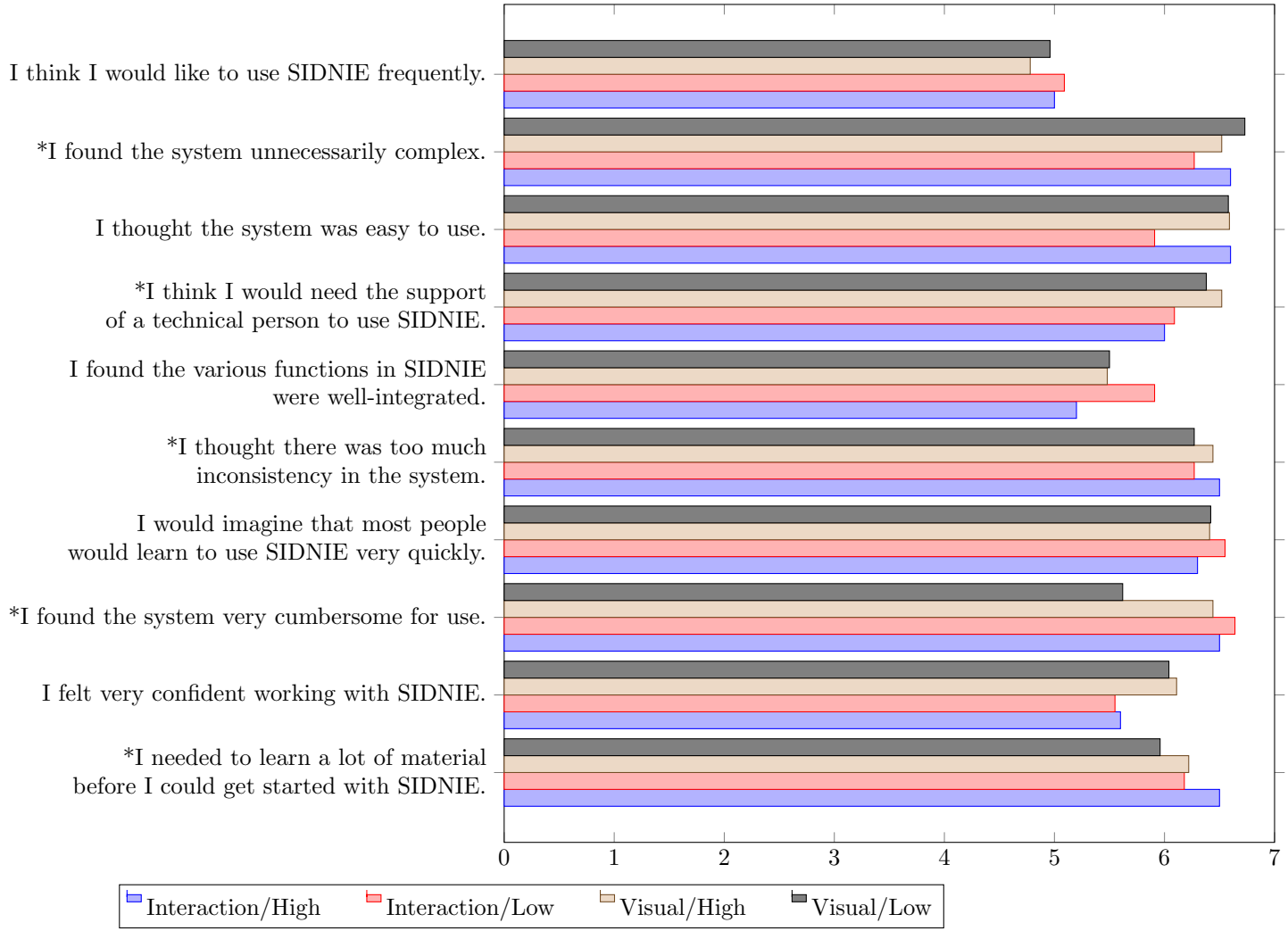


Figure 5.2: Results of the SUS questionnaire separated out by condition and experiment. Scores for questions with a * have been reversed so that in this chart, higher scores always represent better usability.

5.2.3.4 Co-presence

We also asked participants questions about their sense of co-presence with the characters in the virtual environment. 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 low fidelity (clicking questions) 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, did you have a sense of working with the virtual patients?", indicating that the participants did sense that they were collaborating with the virtual patients, to a degree. The lowest rated question in the interaction fidelity study was "To what extent, if at all, were you worried about what the virtual patients thought of your performance?", 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 5.8 for mean scores for each of the co-presence questions divided by experiment and condition.

Table 5.8: Mean ratings on a scale of 1 to 7 for the co-presence questionnaire. Scores for questions with a * were reversed, so that higher scores always indicate greater co-presence. H stands for high fidelity condition, while L stands for low fidelity condition.

		Interaction Fidelity		**Visual Fidelity**	
	Fidelity	**Mean**	**SD**	**Mean**	**SD**
To what extent did you feel embarrassed with respect to what you believed the virtual patients might be thinking of you?	H	1.20	0.23	1.52	0.13
	L	1.45	0.22	1.38	0.13
To what extent, if at all, was working with this system like working with patients in the real world?	H	3.90	0.36	4.33	0.20
	L	4.18	0.34	4.46	0.21
To what extent, if at all, did the virtual patients remind you of someone you've met in the real world?	H	3.40	0.50	2.96	0.27
	L	2.91	0.48	3.42	0.28
To what extent, if at all, did the virtual patients hinder you from carrying out the task?	H	1.70	0.19	1.81	0.16
	L	1.09	0.18	1.77	0.17
To what extent, if at all, did the virtual patients help you in carrying out the task?	H	4.60	0.50	4.63	0.30
	L	4.00	0.48	4.73	0.31
To what extent, if at all, did you have a sense of working with the virtual patients?	H	4.30	0.46	4.74	0.27
	L	4.27	0.44	4.88	0.27
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?	H	3.00	0.41	2.74	0.30
	L	1.73	0.39	2.73	0.30
To what extent, if at all, did you feel that the virtual patients were listening to you?	H	4.30	0.58	4.44	0.33
	L	3.55	0.55	4.38	0.33
I felt that the virtual patients were watching me.	H	3.10	0.46	3.59	0.33
	L	2.73	0.44	3.19	0.34
*The thought that the virtual patients were not real people crossed my mind often.	H	2.60	0.50	2.33	0.31
	L	2.55	0.47	2.81	0.32
The virtual patients appeared to be sentient (conscious and alive) to me.	H	3.20	0.50	3.30	0.28
	L	3.18	0.48	3.04	0.29
*I perceived the virtual patients as only computerized images, not as people.	H	2.50	0.42	2.37	0.28
	L	1.45	0.40	2.04	0.29
How realistic were the virtual patients (for example, how they looked, moved, spoke and interacted with you)?	H	3.60	0.32	3.22	0.23
	L	4.55	0.30	3.65	0.24

5.2.3.5 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.

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".

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 was 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 after every question, while most other participants reported looking at the patients every two or three questions. Twelve 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".

5.2.4 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.

5.3 Modifications to SIDNIE

.

Because of the feedback gathered from our usability studies, we have made several modifications to SIDNIE to improve usability, realism, and learning potential.

5.3.1 Patient Room & Character Behavior

We changed the doctor's office to have paintings on the wall instead of a large mural and improved the lighting in response to complaints that the characters' skin looked strange and the room was too busy. Many participants also felt that the SIDNIE character looked unfriendly, so we modified his appearance to show a more pleasant facial expression and dressed him in Clemson University scrubs. Figure 5.3 shows a screenshot of the new patient room with the new SIDNIE character.

In an attempt to increase the sense of immersion, we decided to only use the SIDNIE system with a large screen television as the display. Additionally, we added animations so that the user feels as if he or she is entering in the room at the beginning of the interview and exiting the room at the end of the interview. When the simulation begins, the nurse sees the outside of a door. The door then opens, and the camera is moved in a linear path from the door to face the patients. When the interview is completed, the camera rotates to face the door and follows a linear path to the door, so that it looks like the user is about to exit the room.

Figure 5.3: Screenshot of the improved doctor's office with SIDNIE.

5.3.2 Changed Scaffolding Levels

In the evaluation of SIDNIE's scaffolding, we found that there was a ceiling effect where repeated exposures to the same SIDNIE scenario lead to memorization, so that by the final scaffolding level the students had memorized the correct answers and only selected them. Additionally, users had a difficult time using the scoring interface. To better provide scaffolded learning opportunities, we changed to three scaffolding levels, all of which focused on learning instead of evaluation of learning. Additionally, for each level of scaffolding, we use a different patient scenario.

5.3.2.1 Level One: Observation

In this level, the student simply observes a patient interview. Sidnie, the virtual nurse, "asks" the questions using text to speech, and the virtual characters answer. Each question Sidnie asks meets all the scoring criteria. At the end of the interview, SIDNIE shows the user a summative feedback chart that lists all the questions he used during the observed interview.

5.3.2.2 Level Two: Guided Interview

This level is the same as level two in the original SIDNIE system. Participants are guided through an interview category by category. SIDNIE gives immediate feedback on each question choice and forces a user to continue asking questions in a category until they select a question that meets the scoring criteria. At the end of the interview, SIDNIE shows the user a chart with summative feedback.

5.3.2.3 Level Three: Practice Interview

This level is the same as level four in the original SIDNIE system. Participants are free to choose the categories and questions and do not receive immediate feedback after each question choice. When the user is ready to finish the interview, he or she selects a button. At the end of the interview, SIDNIE shows the user a chart with summative feedback.

5.3.3 Technical Modifications

Previously, with the hand-rigged virtual characters, shapekeys were used to control the lipsyncing. With the MakeHuman platform providing a full facial rig, we created bone viseme animations, and using Mecanim retargeting in Unity3D, we wrote a controller to apply the bone animations to the characters as they spoke. This meant that shapekeys did not have to be created for each new character, but instead the bone animations could be automatically retargeted to each incoming character.

Additionally, we modified the SIDNIE interface to automatically advance through the scaffolding levels loading in new scenarios for each level instead of requiring a restart for each new scaffolding level. We also re-implemented the entire system in C# (instead of Boo) to increase stability and speed.

Chapter 6

Participatory Design of the Scenario Builder Tool

Participatory design involves all the stakeholders of a product during the design phase [Sharp et al., 2007], in our case those stakeholders being nurse educators and software developers. This iterative, participatory approach is an effective method for interdisciplinary software development, leading to decreased development costs [Rauterberg et al., 1995]. Collaborating with the nurse educators in developing the scenario builder tool increases the satisfaction of the nurse educator regarding the final product [Kujala, 2003], while testing and feedback between each iteration results in a technically sound final product.

6.1 Early Prototype

Since SIDNIE was designed with extensibility in mind, some of the work towards creating a scenario generation tool was already completed. The learning scaffolding structure is independent of the questions, answers, and scorable criteria, so changing the verbal interactions and learning criteria for a patient scenario is as simple as changing the database. The challenges that remained primarily centered around two necessary components of a scenario creation tool: (1) t he implementation of a usable system for virtual character creation so that nurse educators can specify the demographic and physical characteristics of their patients, and (2) the design and implementation of a tool to

input the verbal and nonverbal behaviors as well as the learning criteria into the scenario database.

6.1.1 Specifying a Character's Appearance

Character creation is the most time-consuming part of the process of creating a new scenario by hand. Each requested character must be individually modeled, rigged, and animated. However, advances in the Unity3D engine and contributions to open-source software have the potential to streamline this process, making it possible for nurses to specify their own characters' appearances.

The open source project MakeHuman [MakeHuman, 2013] uses a single base mesh for human geometry, then applies shapekey morphing to the mesh to modify nearly every aspect of the virtual human's appearance through the use of sliders. All the models created by MakeHuman are public domain, so there is no restriction on their use. MakeHuman also provides a high-quality rig for its characters along with shapekeys for visemes (mouth shapes) and a range of facial expressions. Additionally, In its recent 4.0 release, Unity3D implemented a new animation system called Mecanim which automatically retargets animations to new characters. Because the MakeHuman rigging is compatible with the Mecanim system, any animations made for one MakeHuman character will be usable for every MakeHuman character.

We wrote a script that allow a user to specify the salient characteristics of a virtual character, then uses a modified version of the MakeHuman project to generate a character with those demographic characteristics while selecting random values for other appearance values, yielding unique characters each time. The character model is then automatically imported, rigged, and textured in Unity3D at runtime, and can immediately act in place of the patients. See Figures 6.1 and 6.2 for screenshots of generated characters.

6.1.2 Input Interfaces for Scenario Content

Through consultations with nursing collaborators and application of usability principles, we made an initial design concept for the scenario builder tool. Since the "wizard" interface is commonly understood, we developed image-only prototypes in Powerpoint of a wizard for scenario content generation. A breadcrumb trail at the top of the screen enables users to return to any step of the process they have already completed to make modifications.

The first task is character selection. Nurse educators are first encouraged to select a char-

Figure 6.1: Three examples of five year old Caucasian females generated by the first iteration of the character building scripts.

Figure 6.2: Three examples of ten year old African males generated by the first iteration of the character building scripts.

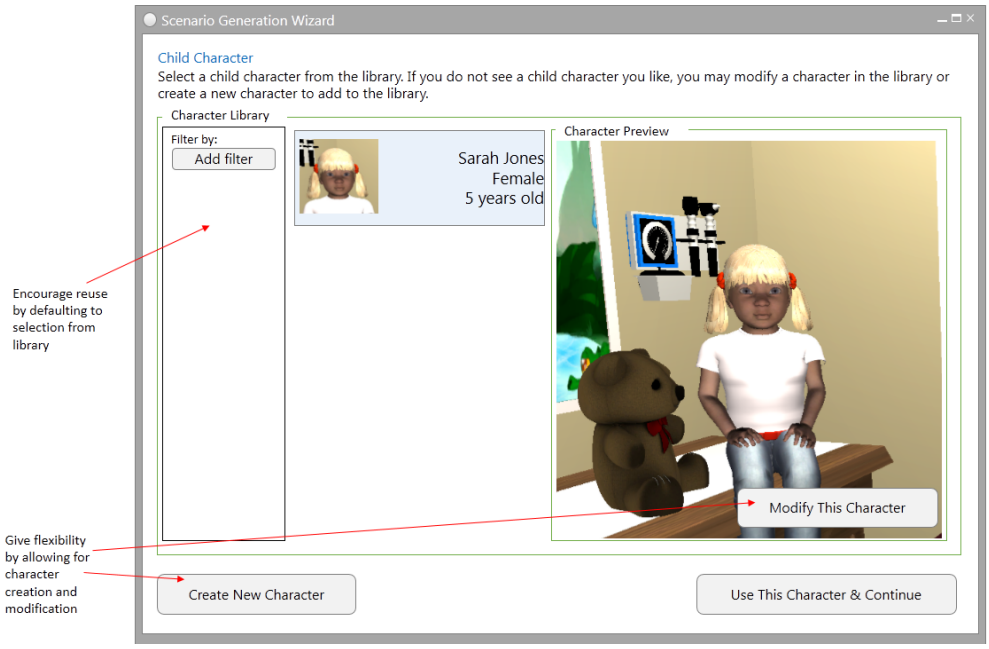


Figure 6.3: Nurse educators will first be encouraged to select a character from a library, encouraging reuse.

acter from the existing library of characters, but there is no suitable character, the nurse educator may edit an existing character or create a new character. The new or edited character is then contributed back to the character library. The nurse will complete this first step for both the child and the accompanying adult. See Figures 6.3 and 6.4 for annotated drawings of a potential interface.

Next, the nurse educator would specify information found in the patient's electronic medical record. The learner would be able to review this information before the interview. Additionally, data from the electronic medical record would be used to automatically generate some questions and answers for the simulation. Similarly, in the next step, the nurse educator would select symptoms that the patient should present. The symptoms selected also help generate questions and answers for the interview, as well as determine idle-state animations for the virtual characters (for example, if the patient has a runny nose, a sniffle animation may play from time to time). See Figures 6.5 and 6.6 for annotated drawings of a potential interface for these two steps.

The nurse then specifies the learning goals for the simulation. In addition to the learning goals specified by Diers [Diers, 2008], the nurse may add learning goals, which will be contributed back to the library for future use. See Figures 6.7 and 6.8.

Using the information the nurse has provided in the preceding steps, the system would then

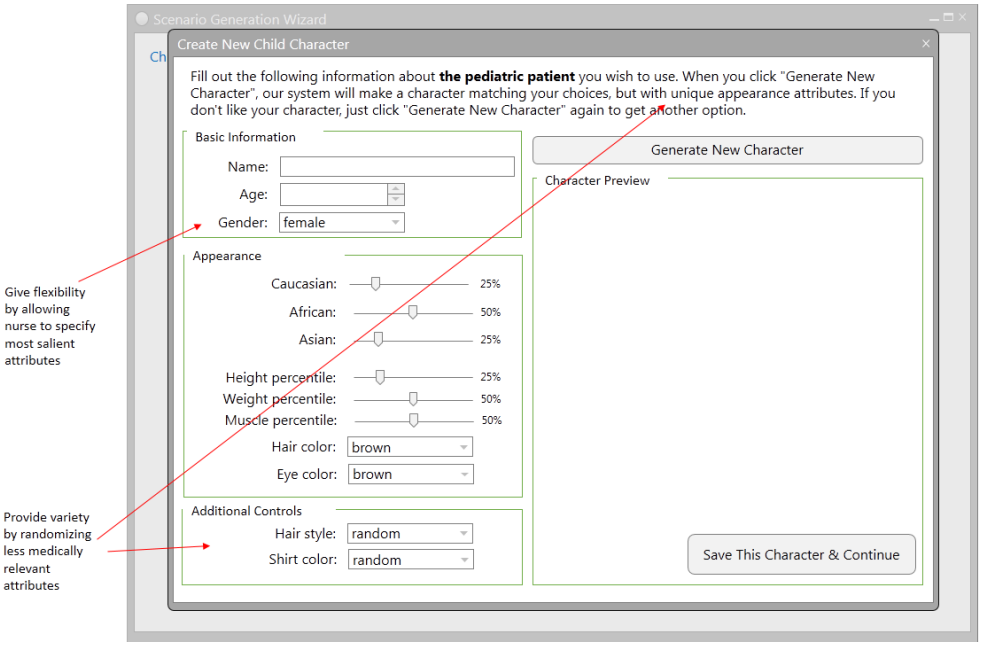


Figure 6.4: If there is no suitable character, the nurse may edit an existing character or create a new one.

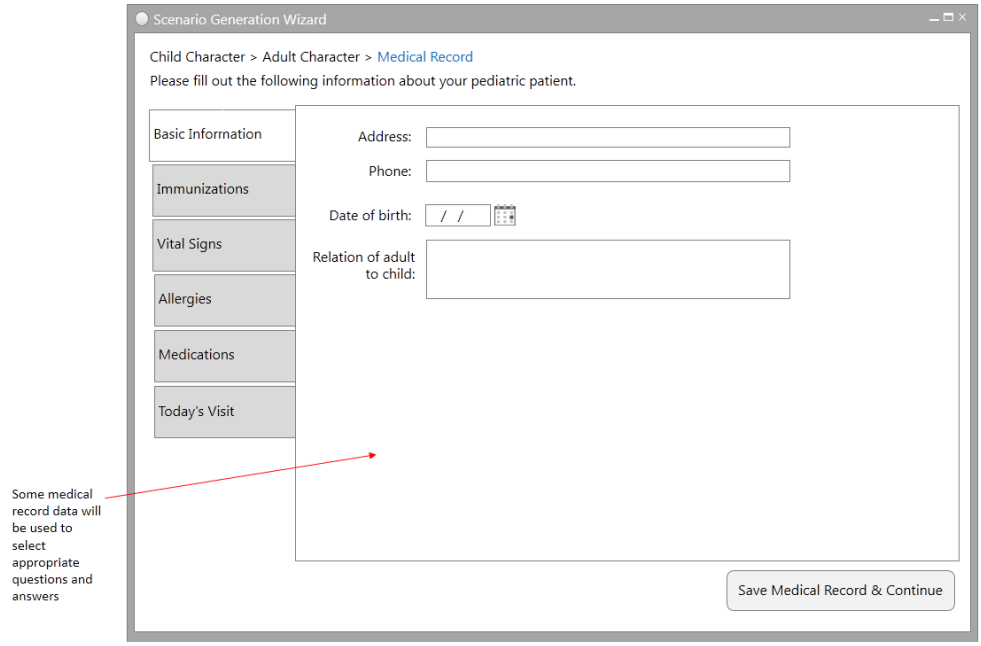


Figure 6.5: The nurse educator fills out the information for the patient's electronic health record.

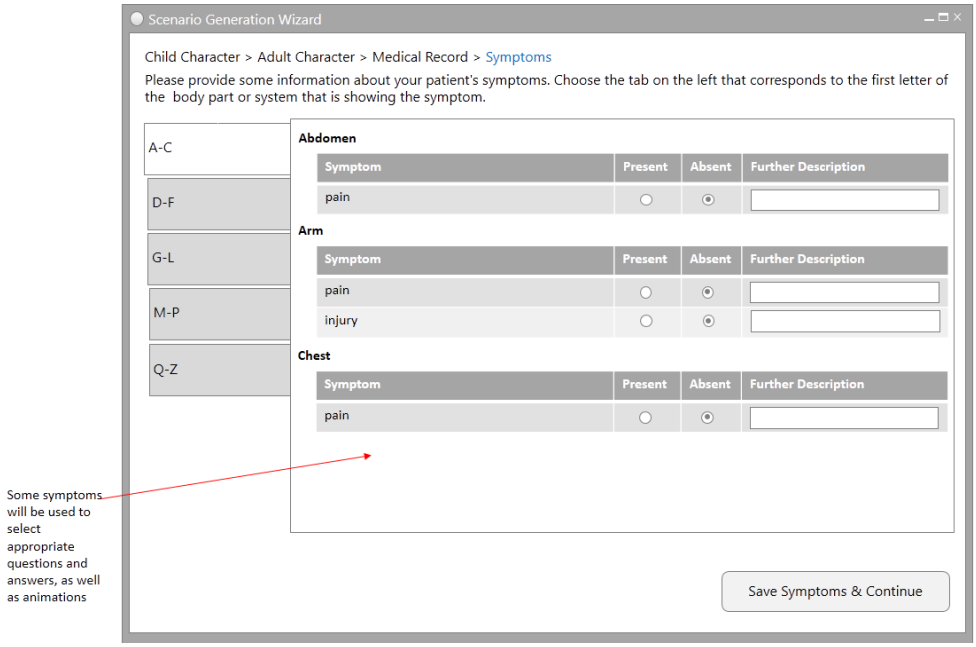


Figure 6.6: The nurse educator checks off the symptoms that the patient should present. This is used to automatically select some questions and answers for the scenario, and to generate animations.

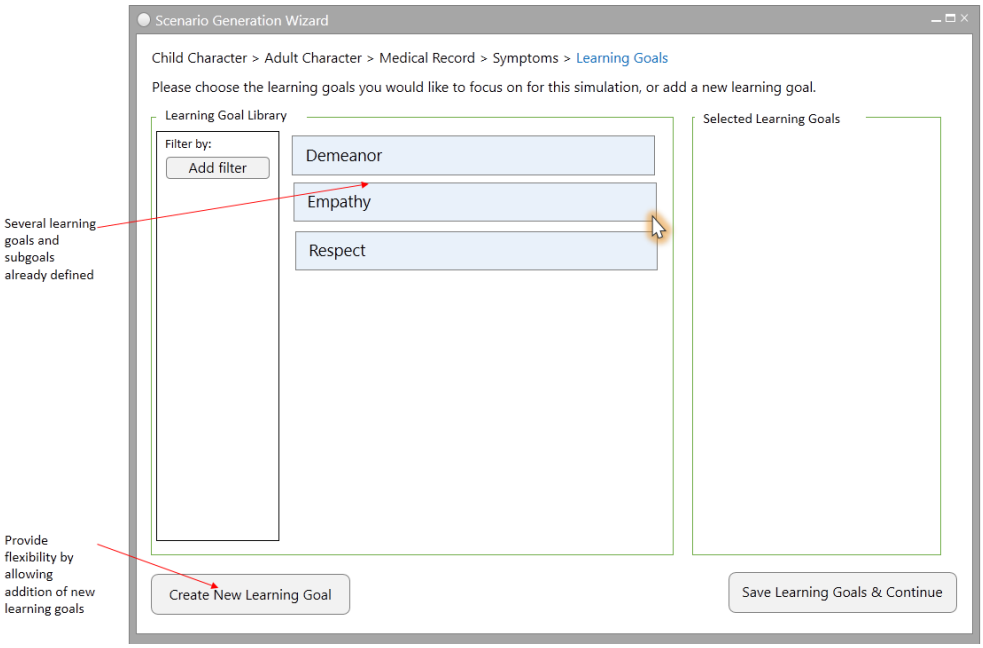


Figure 6.7: The nurse educator selects the learning goals for his or her scenario.

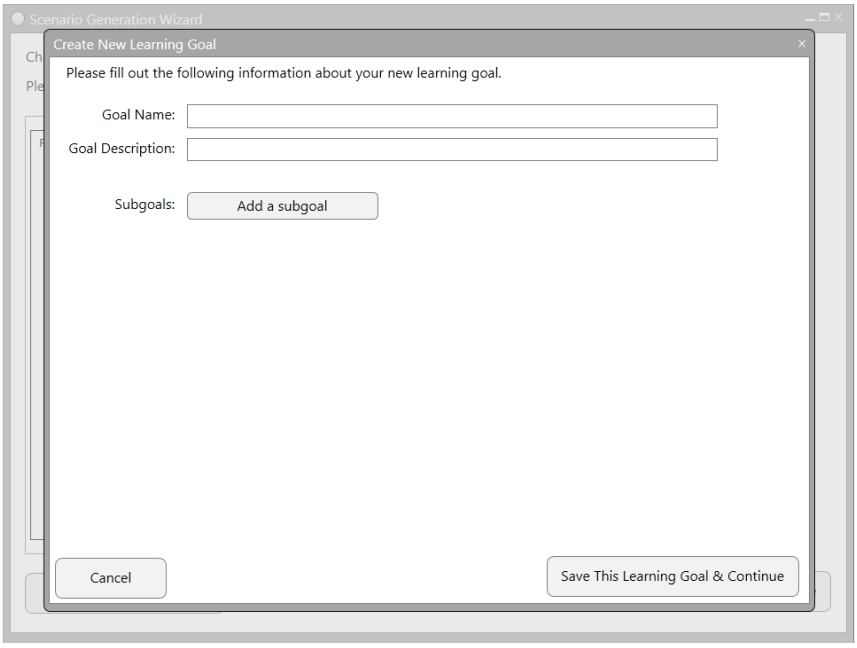


Figure 6.8: The nurse educator may add new learning goals not already present in the system.

automatically select a set of questions and answers that match the patient information and learning goals the nurse provided. This reduces nurse educator burden by providing a starter set of questions and answers that he or she can then modify to suit his or her tastes. The nurse educator can edit or remove any question/answer pair she does not like, and can preview how the characters will answer the question in-place. See Figure 6.9 for a concept for the initial suggestions. If the nurse chooses to add another question, he or she will first be directed to the existing question library to encourage reuse. If he or she still does not see a well-suited question, he or she can add a question and answer. The question will be contributed back to the library for future use. See Figures 6.10 and 6.11.

6.2 Participants and Experimental Design

We recruited three professors from the School of Nursing at Clemson University to help with the participatory design of the scenario builder tool. The design process lasted approximately four months and took place in four sessions.

Due to scheduling constraints with the nursing faculty, we determined that it would be best to conduct as many sessions as possible individually and remotely. The first session would

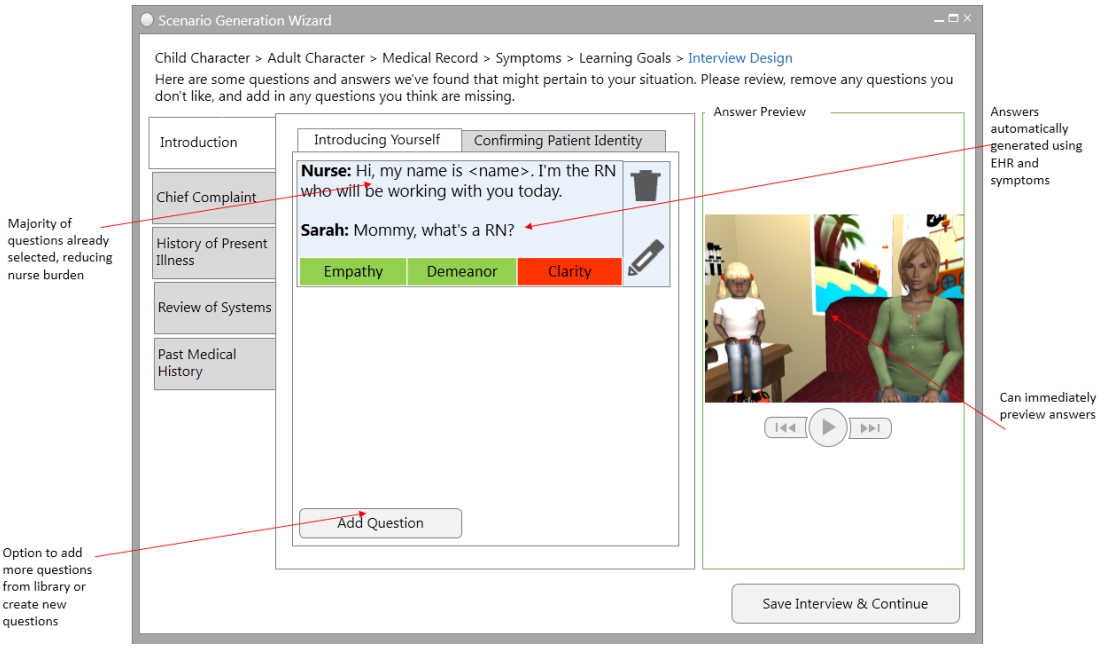


Figure 6.9: The nurse educator can review the questions and answers automatically selected based on her input into the system.

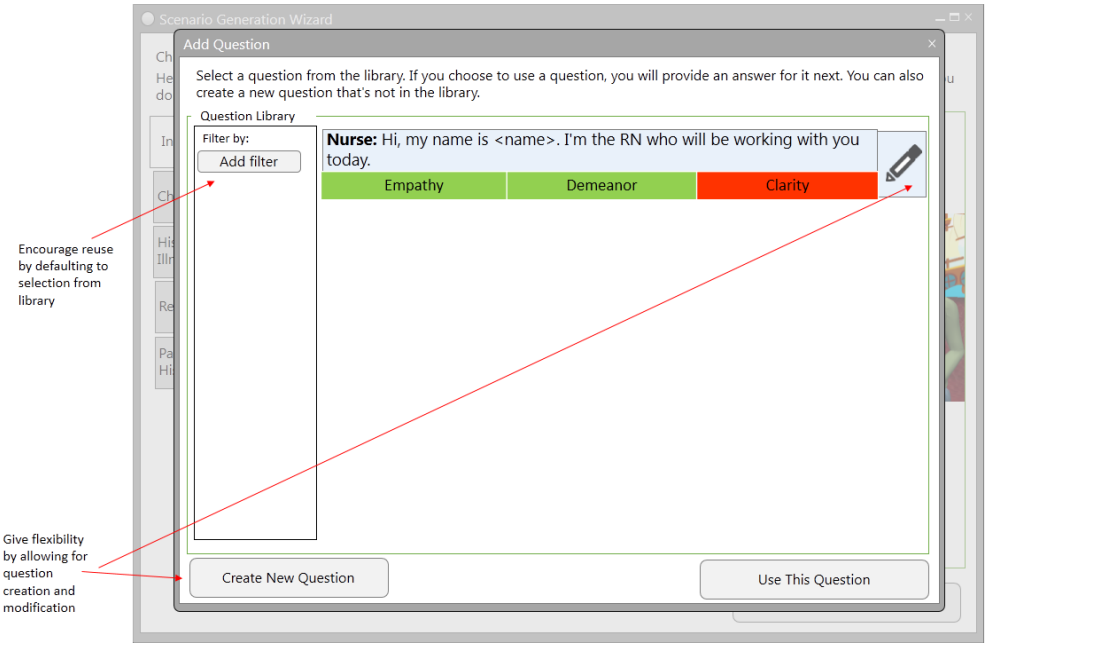


Figure 6.10: If the nurse educator wants to add a question, he or she is first directed to the question library to see if there are any suitable questions already present.

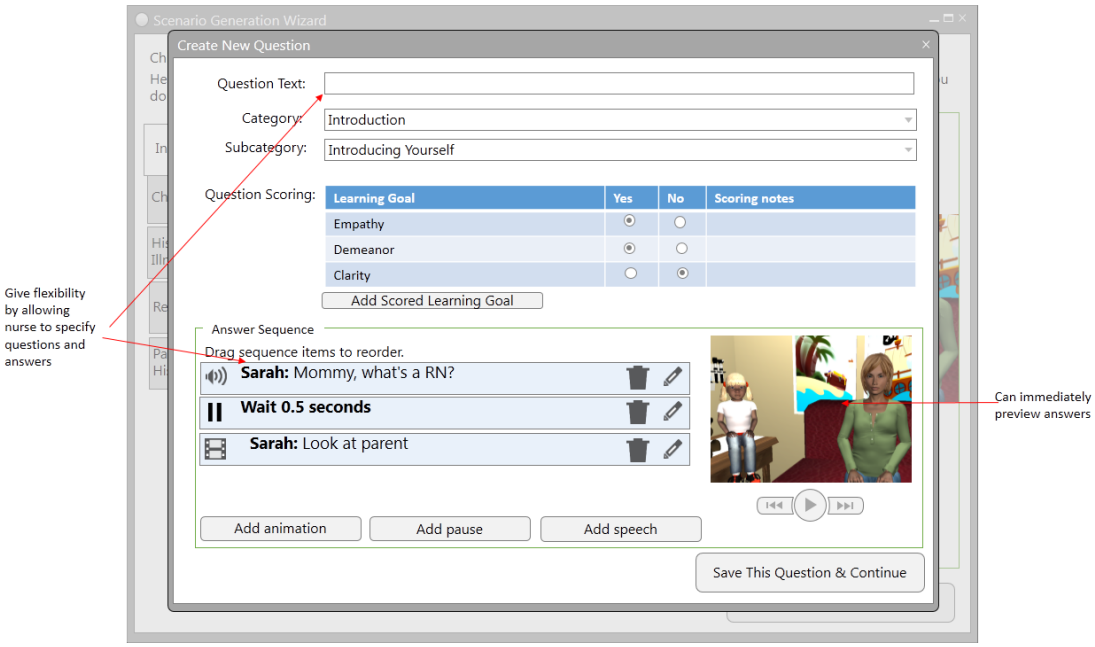


Figure 6.11: The nurse educator may add a new question or edit an existing question to better suit the simulation.

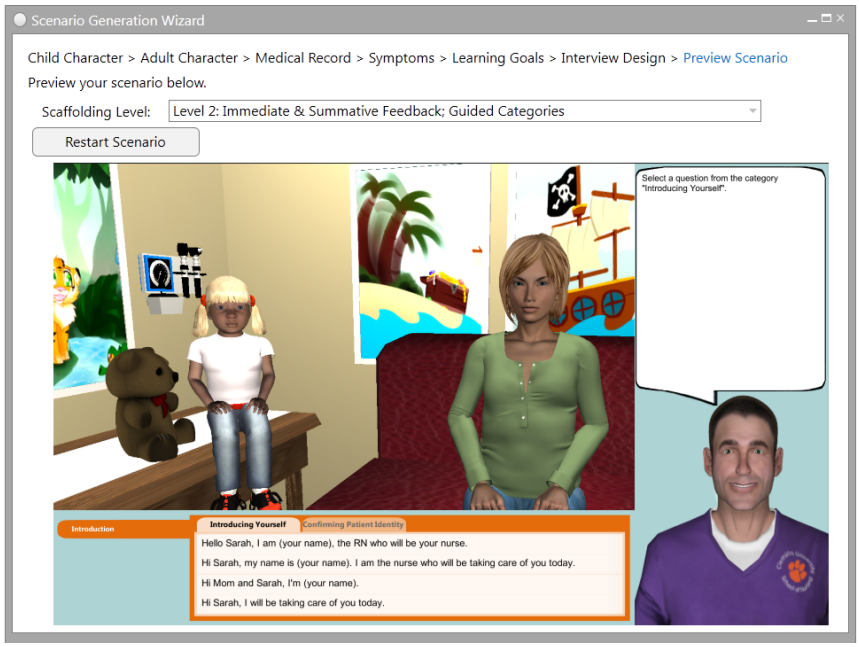


Figure 6.12: After the nurse completes scenario generation, he or she can preview the simulation in each scaffolded learning level before packaging it for student use.

be an informational meeting conducted as a group. The following two sessions would be remote, meaning that we would send them a software prototype and they could interact with it on their own time. Although this sacrifices some experimental control, it made participation convenient, and we believe that the "field study" nature of those two sessions revealed far more about the desired uses of the tool than we could have discovered in a controlled experiment setting. The final session was conducted in person, with me observing the interaction with the prototype, answering questions as they arose, noting any technical errors, and asking questions about usability when the participant appeared confused.

6.3 Session One: Introduction

In the introductory session to the participatory design, we met with all three nurses in a group. We began by collecting their informed consent, then we gave a presentation on the purpose of the project and showed a demonstration of the SIDNIE system. We answered any questions that the participants had about the project.

As a part of the introductory session, we gathered the participant's feedback on how they would like the design process to occur. Although some participatory design processes begin with the participants giving their own ideas for how the product should be designed, our participants expressed that that open of a procedure felt overwhelming to them and that they would prefer to have a starting prototype and then discuss the advantages and disadvantages of the prototype. After the participants expressed this opinion, we showed the participants the early image-only prototype described above, and demonstrated how characters could be created. The participants were most excited about the character generation capabilities and had no further comments about the prototype design.

6.4 Session Two: Initial Prototype

The initial prototype was implemented in Microsoft Blend for Visual Studio, which enabled me to give it a "sketched" visual appearance to remind the participants that not all functionality may be available. Blend for Visual Studio also automatically provided tools where the user could draw on the prototype and enter comments, then send the feedback back to me. See Figure 6.13

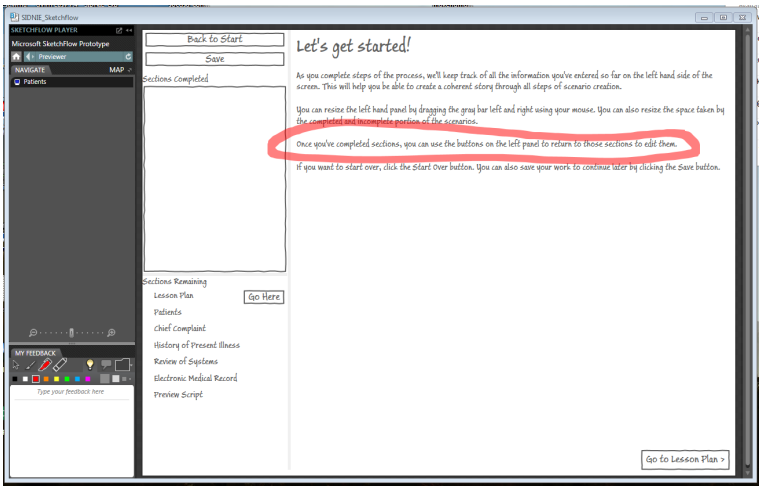


Figure 6.13: This figure shows the prototype running inside the Blend interface. The drawing and commenting tools are located at the bottom left corner of the screen.

for an example of a prototype with drawings.

Because a medical interview is a highly structured process, we used the framework of a typical medical interview to guide our prototype design. My initial design for the prototype was in a "wizard" format, where there was a rigid structure of sequential steps. Participants could not move to the next step until the current step was completed. Additionally, participants could only move backwards one step at a time, and when moving to previous steps in the sequence, all previously stored information in screens beyond that step was removed. Medically, the structure was based on the Medicare guidelines for patient interviews [of Health et al., 2014].

In addition to the wizard interface, the initial prototype relied on "expanders" to make the information on each screen manageable. The user could choose which sets of information to "expand" or "collapse" (make hidden under a header). Also, on the left hand side of the screen, there was an interface to show a summary of the progress so far. That way, users did not have to go back multiple screens to see what they had already filled in.

Each step in the initial prototype is described and pictured below.

6.4.1 Lesson Plan

In addition to creating a medical interview, the nurse educator also needs a way to specify the learning objectives of the simulation for the student and to determine how the student should

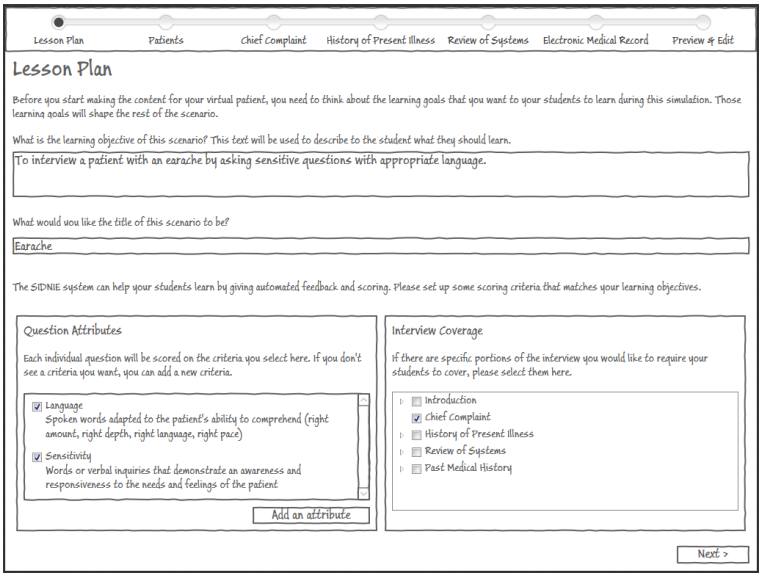


Figure 6.14: Screenshot of the lesson plan screen for the initial prototype.

be scored on their scenario performance. The first screen was designed to get the nurse educator to define the learning objectives ahead of time and to center their scenario creation around those objectives. See Figure 6.14 for a screenshot of the initial lesson plan screen.

6.4.1.1 Scenario Title and Learning Objective

The nurse educator gives the scenario a title, then lists the students' learning objective to be shown in the SIDNIE system.

6.4.1.2 Scoring Criteria

The nurse educator selects the criteria that the interview questions should be scored on. For already-existing criteria in the criteria list, there are questions in the database that are already scored on those criteria. If the nurse educator wants to create a new criteria, he or she could add a criteria here. However, there is no automated scoring process for criteria, so questions would later have to be scored by hand for that criteria for the SIDNIE system to be able to work.

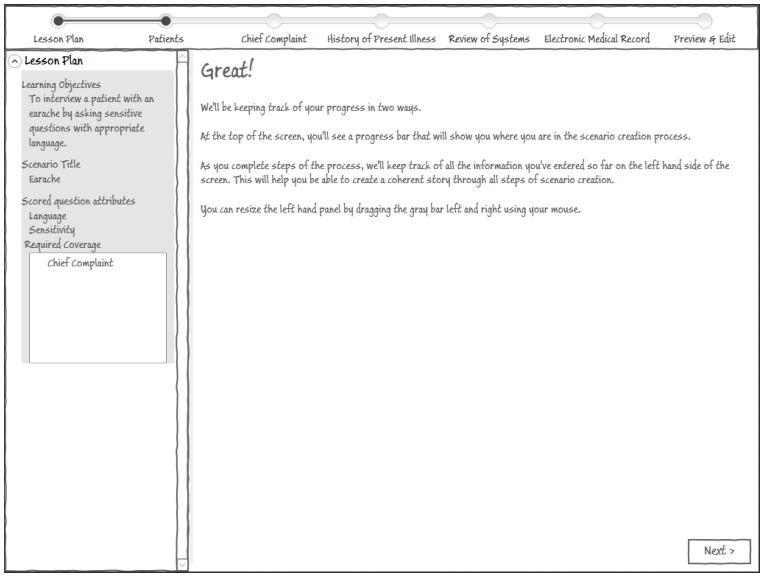


Figure 6.15: Screenshot of the previewer screen for the initial prototype.

6.4.1.3 Required Interview Topics

There are five overarching categories for a typical medical interview: (1) Introduction, (2) Chief Concern, (3) History of Present Illness, (4) Review of Systems, and (5) Medical History. The SIDNIE system is capable of scoring participants for coverage of each of those categories as well as subcategories. In this interface the nurse educator could select which of those categories should be required in an interview.

6.4.2 Previewer

To help the user maintain a sense of context as he or she completed the scenario, after the first step was completed, a bar appeared on the left hand side of the screen that gave a summarized preview of all the steps completed so far. Since there was no way to move to previous screens besides cycling back through them one by one, the previewer gave the user a way to see what was had already been entered without disrupting the flow of the creation wizard. See Figure 6.15 for a screenshot that shows the initial presentation of the previewer feature.

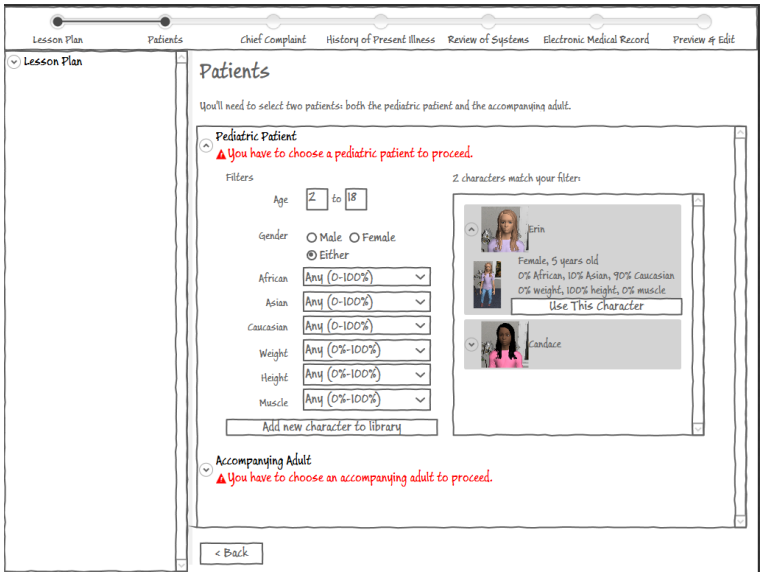


Figure 6.16: Screenshot of the patient selection screen for the initial prototype.

6.4.3 Patients

In the second screen, the nurse educator could select the patients to use for the scenario. The nurse educator had to select both a pediatric patient and an accompanying adult.

There were two options for selecting a patient. First, the user could select a patient from a library of already-created patients. Second, they could choose to create a patient to suit their desired demographics. In this prototype the user could walk through the process of creating a patient but the 3D model was not actually created and displayed, because we wanted to focus the users' attention on refining the information-gathering process instead of drawing attention to the scenario as the end product. See Figure 6.16 for a screenshot of the patient selection screen.

6.4.4 Chief Complaint

In the next step, the nurse educator selected the chief complaint, which is the primary reason that the patient came to the doctor today. There was a list of ten common complaints for pediatric patients, or the nurse educator could choose to create a new chief complaint. See Figure 6.17 for a screenshot of where the user could select the chief complaint.

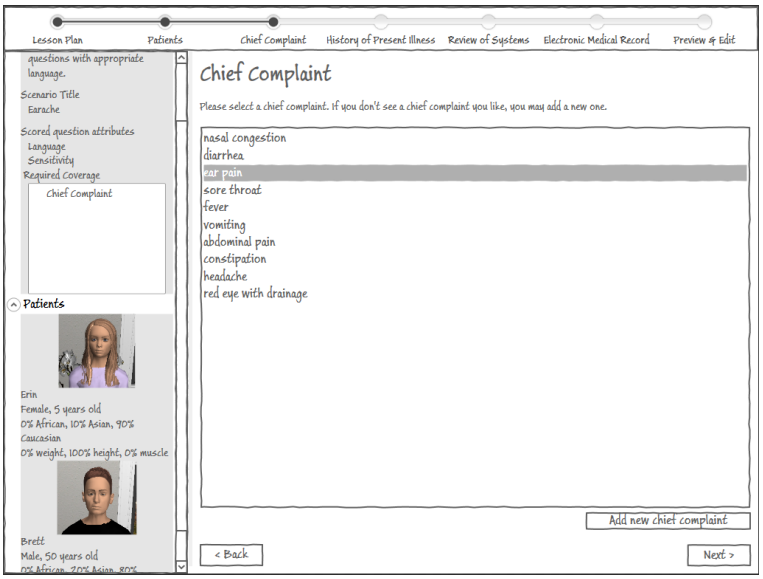


Figure 6.17: Screenshot of the chief complaint screen for the initial prototype.

6.4.5 History of Present Illness

In the history of present illness, the nurse educator provided more information about the chief complaint. My scenario builder tool follows the Medicare-provided structure for evaluating the history of present illness by breaking down the questions into several subcategories. Each subcategory was represented as an expander so that it could be collapsed to reduce the number of items visible on the screen. See Figure 6.18 for a screenshot of the interface for the history of present illness.

Quality The nurse educator was presented with a list of adjectives that could be used to describe any pain associated with the chief complaint (for example, sharp, dull, aching). The options were implemented as a check box interface so that any number of adjectives describing pain quality could be selected. It was also permitted for the nurse educator to select no adjectives to describe the quality of the pain.

Severity The nurse educator had to describe the severity of the pain on a pain scale. To provide animations for our patient to display, we displayed the Face, Legs, Activity, Cry, Consolability (FLACC) pediatric pain scale, which is used to judge pain severity using only nonverbal observation of a pediatric patient. Each subscale has several options to select from, ranging from behaviors that exhibit no pain to behaviors that indicate severe pain. Each subscale was implemented as a set of

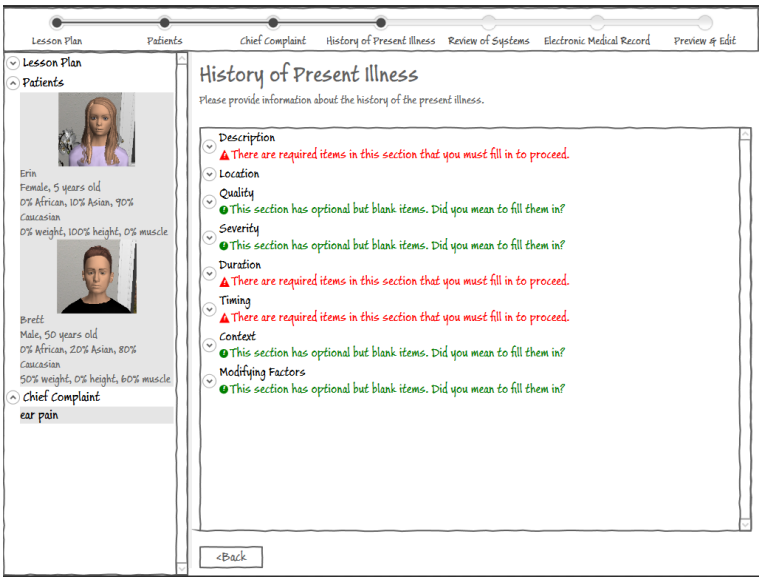


Figure 6.18: Screenshot of the history of present illness screen for the initial prototype. Each component of the history of present illness is represented by an expander with several options below.

check boxes where one or more behavior had to be selected.

In addition to the FLACC scale, we also provided a field where the nurse educator could enter a pain score on a scale of 0 (no pain) to 10 (worst pain imaginable).

Timing The timing of the chief complaint is how frequently the chief complaint bothers the patient. We provided options such as hourly, daily, or constantly. The nurse educatir could also add any other options he or she desired. the options were implemented as a radio button set so that one option must be selected before continuing.

Duration The duration refers to how long the chief complaint has been a concern. Similar to timing, we provided some common options in a radio button format. The nurse educator could also add other options.

Modifying Factors Modifying factors are those situations that make the chief complaint better or worse. The modifying factors were presented as a set of checkboxes, although no modofying factors were required to continue in the wizard.

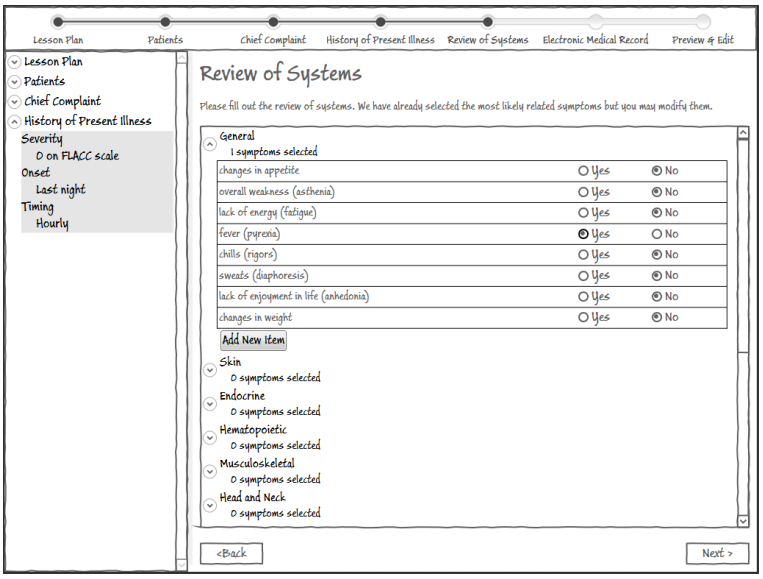


Figure 6.19: Screenshot of the review of systems screen for the initial prototype. Each body system is represented by an expander with several related symptoms that can be selected.

6.4.6 Review of Systems

In the review of systems, the nurse asks the patient about other relevant symptoms. We compiled a list of 110 symptoms and presented them to the nurse in this section of the wizard. Each symptom was classified into its body system. Every body system had its own expander, which when collapsed showed how many symptoms were selected from that system. See Figure 6.19 for a screenshot of the review of systems interface.

6.4.7 Electronic Medical Record

Finally, the nurse educator filled out an electronic medical record for the patient, including demographic information, vital signs, medications, allergies, and vaccinations. Each of these categories was represented by an expander. The electronic medical record was automatically populated with vital signs and and previous immunizations typical for a patient with the specified demographic. See Figure 6.20 for a screenshot of the electronic medical record.

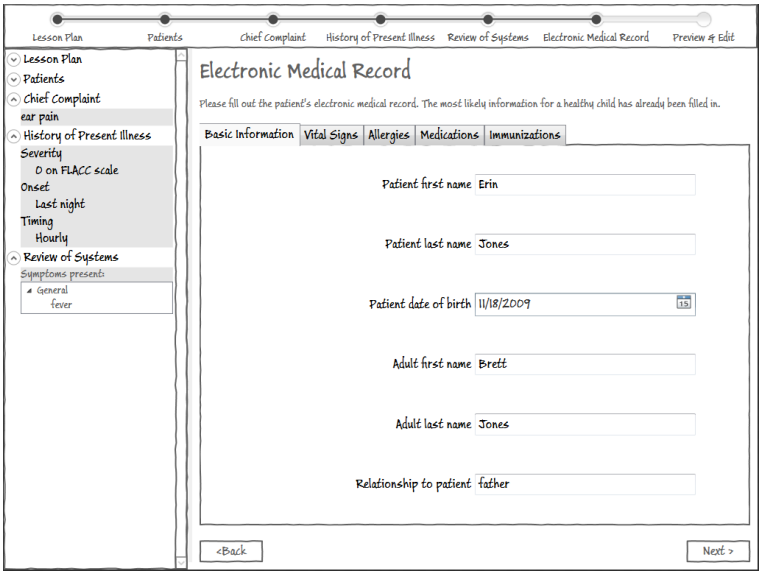


Figure 6.20: Screenshot of the electronic medical record screen for the initial prototype. Each section of the electronic medical record is represented by a tab. Many fields in the electronic medical record were automatically populated according to the patient's demographics.

6.4.8 Evaluation Procedure

After approximately one month from the initial meeting, we e-mailed each participant with a prototype system for scenario generation. We asked them to either create a scenario of their own choosing, or to create a scenario we provided for them for a 7 year old Caucasian female with an earache, accompanied by her father. While they interacted with the scenario, through the Blend prototyping tool, they could choose to draw or write comments on the application that we could later review. We also asked the participants to fill out a questionnaire after they completed interacting with the prototype that asked demographic information and usability questions. See Appendix E for a copy of the questionnaire.

6.4.9 Results

The first prototype, as should be expected, had the greatest number of technical flaws. As the faults were reported, we made note and corrected thm in the next iteration.

None of the participants chose to draw or write comments on the prototype screens, so there were no results to report. However, the questionnaires revealed a great deal of information about how the participants used the system that would not have been discovered in a standard usability

testing environment with a rigidly designed task.

The greatest surprise that the questionnaires revealed was that the nurse educators wanted to use the system in a way that we did not anticipate. While we envisioned the scenario creation as a process that would take place in one sitting with only one pass through the wizard, the nurse educators clearly wanted to use it in a less linear fashion. They requested "hyperlinks to previous pages" and "the ability to edit what I had already put in." The participants expressed frustration when their work was erased when they went back to previous steps. Additionally, there were comments indicating that the nurses wanted to be able to work on their scenario in multiple sittings. One participant wrote when encountering a problem that she would "open [her scenario] up and try again tomorrow." None of these actions were supported in the current prototype, so it was clear that we needed to make some changes to support these use cases.

We also learned a few things about some potential creation processes. One nurse indicated that she pulled up a case study on one of her monitors then tried to make her scenario match that case study. While this did not necessarily affect the design of our system, it did give me a starting point to consider the content of case studies and how that might match up to creating a scenario. Another nurse stated that "more pre-installed scenarios would help...most educators just want to click a link without filling in all the information." This lead me to consider how we could design the system to fill in more of the necessary data points automatically, reducing user workload in creating a scenario.

Although character creation was not fully implemented, two out of three participants listed seeing the character images as their favorite part of scenario creation. (The third participant did not list a specific attribute.) The most salient usability problems were the lack of the ability to return to previous pages, and also some difficulty with the amount of scrolling required on some pages. One participant also suggested the ability to upload images to appear in the patient's electronic medical record.

We also asked five Likert-style questions about using the system. When we asked whether the participants understood what they were supposed to accomplish with the system, all three participants reported high understanding (4 or 5 out of 5, where 5=strongly agree). All three participants rated the question "I thought this prototype was easy to use" at 4 (agree). For the question, "There were a lot of things about this prototype that I did not understand", all participants responded strongly disagree (one participant), or disagree (two participants). For "I enjoyed using

this prototype," participants either agreed (one participant) or strongly agreed(two participants). The widest spread of responses was for the question "I found this prototype frustrating to use." One participant strongly disagreed, one participant disagreed, and one participant agreed.

6.5 Session Three: Refined Prototype

The second prototype iteration was very similar to the first iteration. We kept the "sketch" style provided by Blend to communicate that the work was still unfinished. The evaluation procedure was also similar–the nurses received the prototype by e-mail and were free to work on it during their own time.

6.5.1 Prototype Improvements

We chose to delay resolution of several complaints and requests. First, we decided to still leave out the character creation capabilities since we still wanted the evaluation's focus to be on the patient interview data collection. Second, in response to complaints about problems with scrolling, we made some interface changes in hopes of making interaction less difficult, but we did not reduce the amount of scrolling required. The changes we chose to make are outlined below.

Saving and Loading In response to the nurse educators' desire to work on their scenario through multiple sessions, we implemented the ability to save and load scenarios at any point in the creation process. Loading a scenario would put the user at the same screen that he or she left off at in the creation process

Improved Navigation To enable easier return to previously completed screens, in the sidebar previewer, under each completed section we added a button where the user could return to that section. Although the workflow was still roughly in a "wizard" format, we changed the data model so that already-entered information was not deleted when returning to a previous screen. Once in a previously completed screen, the user could either choose to linearly travel forward through the next screens or could use a button to jump directly to the last screen they completed. See Figure 6.21 for a screenshot of the updated previewer. Since the previewer also listed the sections yet to be completed, we removed the progress bar at the top of the screen.

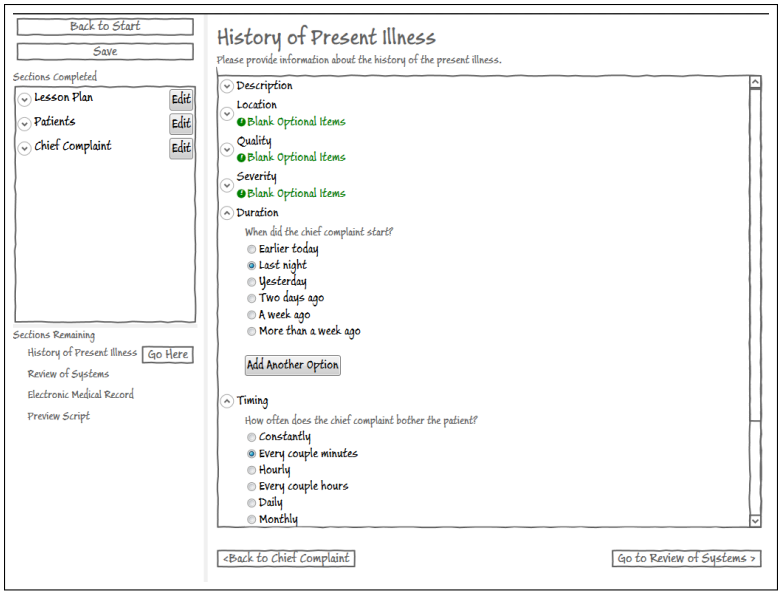


Figure 6.21: This figure shows the history of present illness screen in the second prototype. On the left hand side of the screen, there are buttons to return to previously completed sections, as well as a listing of the sections to come.

Prefilling Information Based on Chief Complaint To reduce scenario creation workload, we implemented a solution to "prefill" certain scenario fields if the nurse educator chose a chief complaint that had already been created by another user. When a chief complaint was first created, the nurse educator was required to fill out a typical review of systems and a typical history of present illness for that chief complaint. Then, when using the chief complaint in future scenarios, all of the originally collected data would be loaded and the nurse educator could then just tailor it to his or her specific scenario (or choose to leave it unaltered).

Previewing Autogenerated Questions Since many of the content collection issues were resolved, we decided to implement the ability to preview the questions that were automatically generated. We created a database of question and answer "templates" for each category in the typical medical interview. Each question template was scored for the two pre-created scoring criteria. Within each question were placeholders for content that the nurse educator would select within the scenario builder tool. Once the nurse educator input all the necessary data, we used database queries to select questions for each category that met all of the selected scoring critieria, met each of the scoring critieria individually, and met none of the scoring criteria. The questions were then pro-

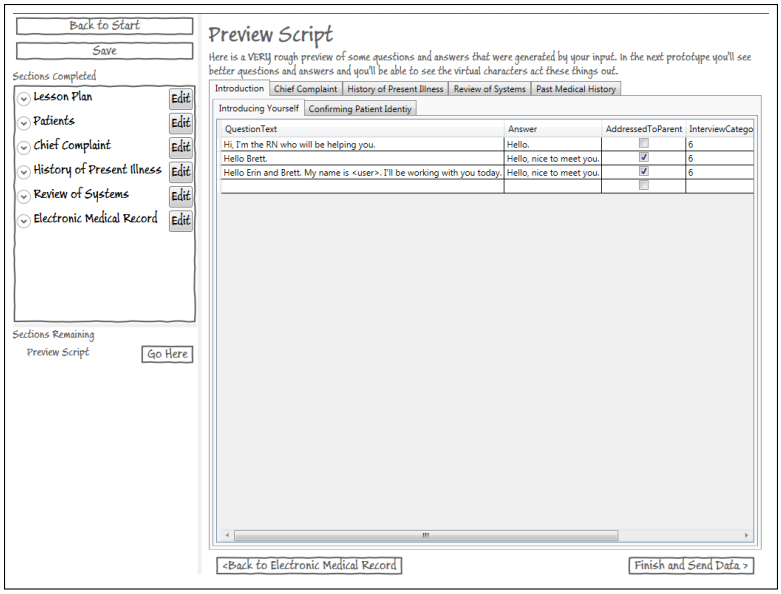


Figure 6.22: A screenshot of the preview screen for the questions automatically generated using the information in the scenario builder.

cessed to remove the placeholders and put the scenario content in its place. The generated questions and answers are not necessarily high enough quality for immediate use, and are not tailored to any specific patient demographic. The questions and answers are intended as a starting dialogue that is consistent with the medical portion of scenario, which can then be modified to best fit the patient demographic. This reduces the burden of scenario generation by making it a task of customizing questions and answers instead of having to come up with them each time.

The selected questions and answers were then shown to the user in a data table format. The user could edit the question text as well as alter the question scoring. For this iteration we chose to not give the user the ability to add or remove questions since we primarily wanted to gather feedback on the automatically generated questions. See Figure 6.22 for a screenshot of the question preview.

Expanded Electronic Medical Record In response to user requests, we added another tab in the electronic medical record where the user could insert an image or a text-based note to appear in the electronic medical record when the student interacted with the system.

6.5.2 Evaluation Procedure

Approximately one month after the second session, we e-mailed the participants the new prototype and links to a questionnaire. We also encouraged them to draw on the prototype using the Blend tools if it was relevant. Finally, we modified the program so that completed scenarios would be automatically e-mailed to me once the participant reached the final screen in the prototype. For the task, we prompted the participant to create a scenario with a 7 year old Caucasian female with pink eye and a fever, or any other scenario of their own choosing,

6.5.3 Results

The format of the results was somewhat different for this prototype. Two out of the three nurses chose to draw and add comments on their scenario screens this time through the Blend interface. We was also able to collect completed scenarios through the automatically sent e-mails for two out of the three nurses. Additionally, all three nurses completed the questionnaire after interacting with the system.

When we reviewed the added drawing and comments, we found that this interface was solely used to indicate places where the nurses had technical problems. Each nurse had circled specific buttons or controls and written comments such as, "When I clicked this button, the prototype shut down and I had to reopen it." These comments were helpful in debugging a few remaining problems for the next prototype.

We was able to load in the two completed scenarios. Although we could not gather much information through the scenarios, it provided technical verification that the saving and loading procedures worked. We also noticed that both completed scenarios matched the scenario description we had originally given the participants, and that neither participant changed any of the default information that was automatically filled in.

Finally, we considered the questionnaire results. We asked similar usability and Likert-type questions as in the first prototype. Participants reported fewer technical errors in this questionnaire and began to report more medical and usability errors. Instead of a numerical text entry for the pain scale, one participant commented that it needed to be changed to a pediatric pain scale with faces. One participant requested the ability to elaborate more on the chief complaint in the patient's words. When previewing the automatically generated questions, one user complained that it was

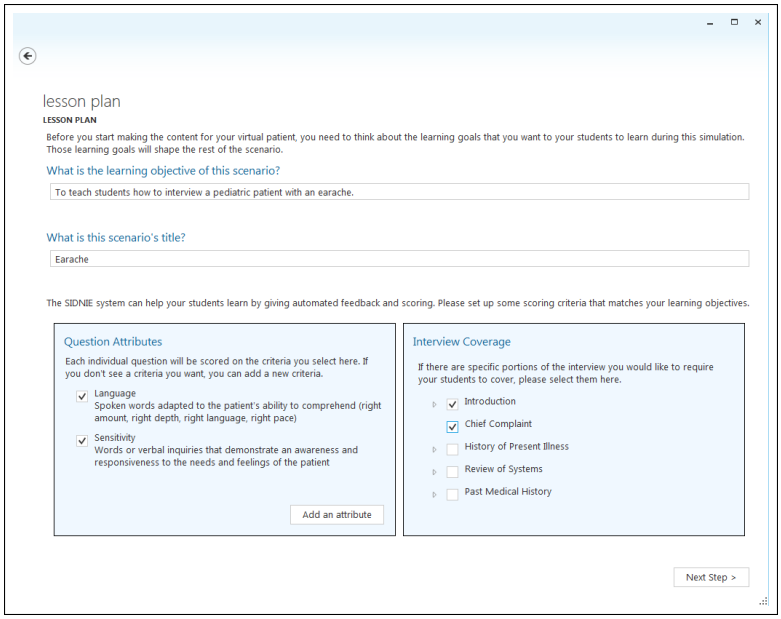


Figure 6.23: A screenshot of the lesson plan in the new Modern UI styled interface.

difficult to read in the chart format.

The problems with navigation between phases of the scenario seemed to be resolved–two out of three reported that tool guided them well through creating a scenario, and there were no reports of problems appearing in the first prototype related to moving between screens or saving and loading the scenario.

6.6 Session Four: Final Prototype

Since most techincal and usability errors seemed to be resolved, for the final prototype iteration we reimplemented our work in a standard windows Presentation Format (WPF) project. We used the ModernUI toolkit [Software, 2015] to give the prototype a modern appearance. Figure 6.23 shows a screenshot of the lesson plan screen with the ModernUI styles applied. With the Blend drawing interface removed, we gained screen space to work with, so as we reimplemented we continued to consider how to best use screen space to increase usability.

Although most scenario creation tasks had been tested in the previous two prototypes, we had delayed discussion of a couple related features until this prototype due to their complexity. For the already created scoring attributes (language and sensitivity), all templated questions in the

database were already scored for the criteria, and there were enough scored questions to provide variety across scenarios. When a new criteria is created, no questions are scored for that criteria. Even if all questions are manually scored, it could be that there are insufficient questions to provide variety in scoring combinations, since it is important to give the student the ability to get the question wrong as well as to get the question correct. This shows there is a need to be able to add questions. Adding a layer of complexity, to maximize reuse, it would be beneficial if users could contribute their new scored questions back to the database as templated questions, so that future users would not have to repeat the process. However, templated questions are somewhat abstract and it is difficult to know how to best create an interface that would support creation and contribution of these questions. In this final session with the prototype, our goal was to discuss this need with each user in an attempt to create a usable interface for future use in the scenario builder.

6.6.1 Prototype Changes

Primarily, the changes in the third prototype were designed to correct usability issues as well as to leverage the additional screen space provided by removing the Blend prototype interface. Additionally, the prototype was tailored for a time-limited and observed session. We disabled the ability to save and load a scenario, and we removed any data validation constraints to simplify the workflow for observation.

Reformatted Scenario Progress Bar To improve navigational capabilities and to streamline the scenario creation process, we removed the side panel previewer and used a menu and submenu system at the top of the screen. This both displayed progress, as only completed sections were displayed at the top of the screen, and allowed for quick navigation back to previously completed sections since a user only had to click on the name of the section and subsection to return there. Removing the previewer panel also freed up screen space for better use. Additionally, a "back" button at the top of the screen provided a standardized method to move back one step. See Figure 6.24 for an example of the top navigation progress bar.

Reducing Scrolling and Unified Navigation In the previous prototypes at least one user had complained of difficulties with the amount of scrolling needed for each screen. This was in part due to the prototype's reliance on "expanders" to separate out different pieces of the interview into

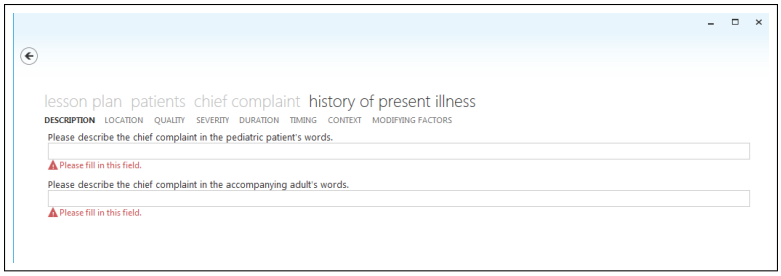


Figure 6.24: A screenshot of the top navigation bar with menus (large text) and submenus (smaller text).

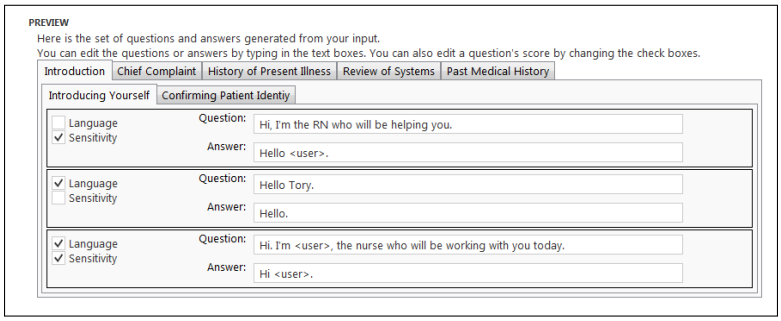


Figure 6.25: A screenshot of the preview screen for the questions automatically generated using the information provided in the scenario builder. The interface is improved to make it easier to read.

manageable chunks. With the new menu and submenu paradigm, we chose to remove the expanders and move each section of the interview into a submenu page of its own. This reduced the amount of scrolling greatly. Additionally, we translated the tabbed interface of the electronic medical record to a page-by-page interface to keep consistency.

Improved Previewing of Generated Questions In the second prototype, users had difficulty reading the previewed questions in the table format. To remedy this usability issue, we changed the interface to make the questions more readable and the scoring interface more accessible. See Figure 6.25 for a screenshot of the improved interface.

Character Creation For the final prototype, we added in the capability to create a character. The user is first presented with a "library" of characters. If he or she does not see a suitable character, the user can click a button to create the character. The user then specifies the character's demographics information (see Figure 6.26 for a screenshot), and selects the "Create this Character" button.

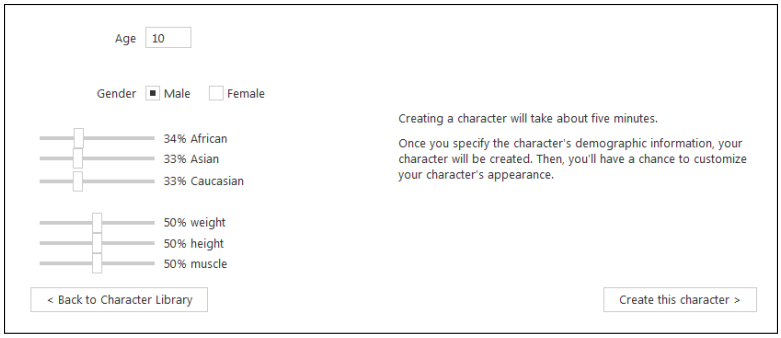


Figure 6.26: This interface enables the user to select the demographic characteristics of the character they want to generate.

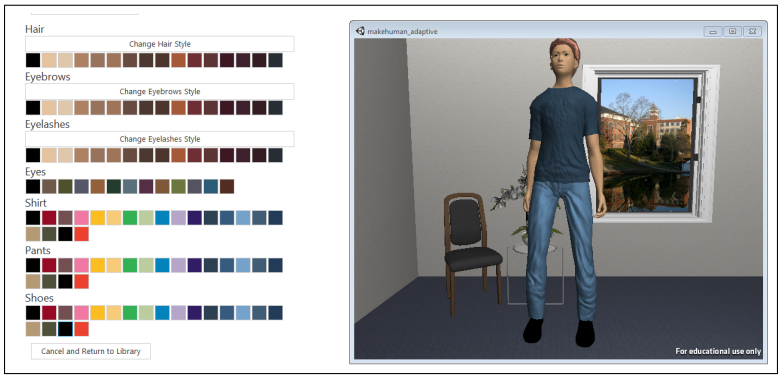


Figure 6.27: This interface enables the user to select more fine-grained appearance characteristics of their character, including hair and clothing colors, and hair and eyebrow styles.

After approximately 90 seconds (the time it takes for MakeHuman to generate the character), the interface then changes so that the user can customize their character by selecting styles for hair, eyebrows, and eyelashes, and colors for hair, eyes, and clothing. The character is loaded in to a Unity3D window so that the user can see how their customizations affect the character's appearance in realtime. See Figure 6.27 for a screenshot of the interface for customizing a character.

Finally, the user can choose to add the patient to the library, and can then select their created patient to use for the scenario.

6.6.2 Evaluation Procedure

The procedure for this evaluation was different from the previous iterations. We scheduled each participant individually for a one hour session. We also scheduled each session two weeks apart, which gave me time to implement suggested changes between participants. This resulted in three

rapid prototype iterations, each informed by one nurse educator, instead of one overall iteration guided by all three nurse educators. We did not use any questionnaires to collect data but instead observed their use of the tool, took notes as they encountered errors, and talked with them about any difficulties in usability that they encountered as they occurred.

6.6.3 Participant 1 Interaction

The first participant spent approximately one hour interacting with the system and discussing the results with me. This participant chose to talk through most of what he was thinking and doing while he interacted with the system.

In the lesson plan, the participant chose to make a scenario about a teenaged girl whose chief complaint was a lump in her breast, since that scenario was one that he was planning to cover in class at the time. The participant selected both the preset scoring criteria for language and sensitivity and opted to not create any other critera because he wanted to "not make it too complicated". He checked off all sections and subsections of the interview to be required.

The participant chose to create a teenaged female character for his scenario. Due to a technical oversight, the character was not loaded in correctly. We was able to correct the problem and the character was then loaded in appropriately. The participant was able to customize the character although he had some difficulty in selecting the hair, eyebrows, and eyelashes styles–the button text that said "Change Hair Style" was unclear to him. Additionally, the participant wanted to be able to see the patient's face close up as he adjusted the eye color, eyebrows, and eyelashes, and suggested that we enable that in the future, especially for fine-tuning character appearances in scenarios where there may be a "cultural component" to the scenario's educational goals. Finally, once the participant finished creating the patient and was returned to the library screen, he believed that he had already selected the patient and should proceed to the next step, while in actuality he had only added the patient to the library and needed to select it. We noted this as a potential usability flaw. Next, the participant selected an accompanying adult from the patient library.

The participant created a new chief complaint, which he titled "breast lump". The chief complaint creation dialogue prompted him to fill out the review of systems for a typical patient (see Figure 6.28). He began to check off symptoms such as anxiety and change in appetite, saying "she's probably been worried about this for a while". We then stopped to ask him whether he was filling out the review of systems for his particular patient, to which he answered that he was. We

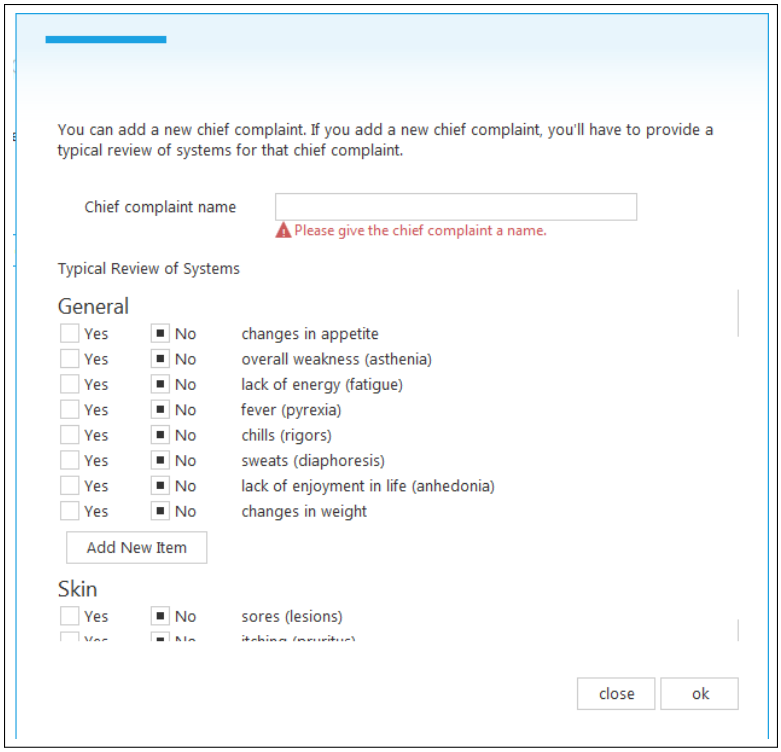


Figure 6.28: This dialogue enables the user to create a new chief complaint and to enter the typical review of systems for that chief complaint.

explained the instructions that were written on-screen–that this review of systems was intended to be symptoms that every patient with this chief complaint exhibited, and that he could customize the review of systems for this specific patient later–and he understood and complied to the instructions. However, we noted this as a potential usability flaw of the system.

The participant filled out the history of present illness and review of systems without problems. He had several comments about the electronic medical record. For the vital signs, although a typical value was automatically selected, he suggested that the typical range of values would also show up next to the box so that the user would not have to look up the range if he or she wanted to modify it to something atypical. When trying to add a medication, immunization, or allergy, he did not understand the "blank line" cue provided (see Figure 6.29) for adding a new item to the table. Finally, for the medication table, he suggested to add fields for route, frequency, and indications in addition to the existing field of medication name, dosage, and reason. The participant also successfully added a supplemental image to the record.

Finally, the participant entered the question preview screen, where several usability problems

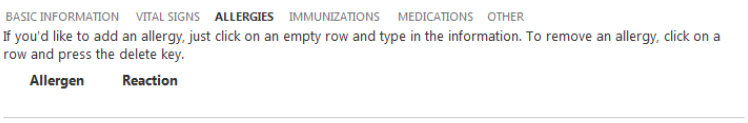


Figure 6.29: This is the electronic medical record's interface for the patient's allergies. To enter a new allergy, the user could click on the gray line under the headers and type the allergen and the corresponding reaction.

were evident. First, instead of using the tabbed interface to navigate through the question previewer, he wanted to use the "Next" button to go through each section. The tabbed interface violated the consistency that was present in the rest of the scenario creator, where consecutive pages could be navigated to through the "Next" button. The participant also pointed out that in the question and answer interface, there was no indication as to whether the parent or the child were speaking the answer. Finally, the participant started deleting all question alternatives except for the ones that he thought were correct according to his criteria. We asked him why he was deleting the questions and he responded that they were not good questions. Although there was direction text at the top of the screen that explained that these questions were alternatives that would be presented to the student, he had not read the direction text and immediately assumed that he was supposed to narrow it down to only the good questions. Once we explained the purpose of the question alternatives to him, he understood, and he tailored the questions and the answers to his specific case.

The participant was overwhelmed by the number of questions available, especially in the review of systems. For two scored criteria, the question selection algorithm we used selected one question that was correct for both criteria, one question that was incorrect for both criteria, and one question that was incorrect for one criteria and correct for the other criteria. So, for 110 symptoms at four questions each, that led to 440 questions. The participant suggested that we narrow down the questions somehow, possibly by excluding questions from irrelevant systems, or giving the user the option to exclude questions on irrelevant symptoms, but was unsure of the best interface to use for that.

Since the participant opted to not create an addition scoring criteria, at the end of the interaction, we asked him how he thought an additional criteria and its corresponding scoring (and any additional questions) should be handled. He understood the problem and agreed that it was a complicated task, but did not have any suggestions for how to best complete it.

6.6.4 Changes After Participant 1 Interaction

We corrected all technical errors encountered during interaction with the participant, and corrected some usability errors. Since this was only one data point, as in previous prototypes, we chose to leave in the majority of the reported issues to see if consensus emerged with the other participants as well.

Addressed Concerns We added in the requested fields for medications within the electronic medical record, and also the normal ranges for the vital signs. Additionally, from our observation, clicking through all the screens seemed tedious and very few screens used the entire available screen space, so we combined several screens in the history of present illness, review of systems, and electronic medical record to take advantage of the screen space and reduce the number of steps necessary to complete the scenario. We reworded several instructions where the participant seemed unsure of what to do, and we reformatted the question previewer so that it was clear who was speaking in the answer.

Unaddressed Concerns We left the character creation interface as it stood in order to gauge its usability better by observing any issues the next participant would have. Similarly, we left the "blank line" interface for adding allergens, medications, and immunizations the same. Since the patient was unsure of how to narrow down the questions presented in the previewer, we left the question selection algorithm intact and made a note to discuss it with the next participant. We also did not change the interface for adding a chief complaint but instead rephrased the instructions in hopes that it made the process clearer.

6.6.5 Participant 2 Interaction

The second participant spent approximately one hour interacting with the system. Although she did not talk through her thought process like the first participant did, she periodically stopped and asked questions as she completed the task.

In the lesson plan, the participant chose to make her scenario about a teenage boy with a sore throat. She pulled up a case study online to use as reference. She selected both of the available scoring criteria and also selected every category and subcategory of the interview as required. She created the pediatric patient and encountered the same usability issues as the first participant. She

wanted to be able to see the character's face better, and was unsure of how to change the hairstyle using the buttons provided. Since both participants had these problems, we determined to change the interface for the next participant.

Similarly, in the interface for creating a new chief complaint, the second participant also started filling out the review of systems for her specific patient. At this point, just as with the first participant, we stopped and explained to her that she should be filling out the review of systems for the typical patient with this chief complaint, and then asked how we could explain that better, since both participants had the same difficulty. The participant responded that "I have my patient in my head, I want to finish my patient then decide what to contribute [back to the database of stored scenario information]". This comment shows that perhaps contribution to the database should be separated out from the wizard interface, since the wizard format is oriented around completing a single scenario, while contribution to the database is based on generalization.

In the history of present illness screen, the participant chose to create several new options and suggested that when an option is created, it should be automatically selected. (This again showed a division between the scenario creation task and database contribution, since a symptom or characteristic contributed to the database would not necessarily be present in the current patient, but a symptom or characteristic would only need to be created in the context of a scenario if it was present in the current patient.)

The patient had no comments on the review of systems screen. When working with the electronic medical record, like the first participant, the participant did not understand the "blank line" interface for adding allergens, medications, and immunizations. She suggested a button and dialogue system for adding items instead.

The second participant was the most familiar with the SIDNIE system, so in the question previewer screen, she understood that the questions and answers presented were the alternatives that students could select from. She suggested a button for adding new questions in addition to the premade questions.

At that point we asked her if she had any ideas about how to add new questions and scoring criteria to the database. Again, enforcing the same theme, the participant responded that she was focused on her current scenario and did not want to worry about contributing to the database at the same time as developing her scenario. We continued discussing this theme with her, asking questions to determine how to best implement this idea. Due to the wizard format, it seemed that users were

approaching the application with a single scenario in mind. If we simply delayed contribution to the database until after the scenario was created, it seemed that authors may not choose to contribute back to the database since their task was finished. Through our discussion we together determined that it might be good to have a separate application or "mode" for database contribution, and to outsource all contribution related tasks to that interface. The participant pointed out that while only nursing instructors would be qualified to create a full scenario to meet learning goals, many other individuals have the skills and knowledge to contribute to the database. For example, she suggested that her undergraduate assistant could easily create and score questions and answers, and could research the typical symptoms present for chief complaints and enter them in. This approach would allow nurse educators to focus on the scenario creation task they are uniquely qualified to complete while allowing less specialized portions of work to be completed by others. Additionally, it would make the wizard interface more streamlined and clear. Finally, since a nurse educator could not use a chief complaint or scoring criteria that had not already been contributed to the database, he or she would be forced to contribute any unique content to the database.

We also asked the participant about the review of systems and how to reduce the number of questions generated, or to eliminate unnecessary symptoms. She suggested that in the database contribution interface, instead of the user filling out the typical review of systems with each symptom marked as present or absent, he or she could mark each symptom as present, typically absent but relevant, and typically absent and irrelevant. The symptoms marked as typically absent and irrelevant could be safely excluded from the question generation by default, while the relevant symptoms could be included by default. In the wizard interface, the nurse educator would have the option to add in the symptoms excluded by default as well as to remove the symptoms included by default.

6.6.6 Changes After Participant 2 Interaction

The primary change due to the interaction with this participant was the separation of scenario authorship and database contribution tasks. We removed the generalized database contribution tasks from the scenario builder wizard including adding a scoring criteria, adding a chief complaint, and creating a typical review of systems for a chief complaint. Additionally, we planned to implement the templated question creation and scoring in the editor interface. In the wizard interface, we left the ability to add symptoms and to add field options for the history of present illness, since

those tasks are low effort and could be necessary in creating a scenario if the presentation of the chief complaint was unusual.

Because the target user of the database contribution is not necessarily a nurse educator, for the changes made between this participant and the final participant we decided to focus on improving the wizard interface instead of spending time implementing the database contribution interface.

State-Aware Prompts To better bridge the gap between the screens, we changed all the generic prompts within the wizard screens to reflect previously selected information. For example, instead of "Please describe the chief complaint in the pediatric patient's words," if the patient's name was Erin and the chief complaint was an earache, the prompt would say "Please describe the earache in Erin's words." This change helped to preserve continuity throughout the scenario.

Streamlined Review of Systems We modified the review of systems interface to indicate the new review of systems categorizations (see Figure 6.30). Symptoms that were relevant were shown at the top of the page as "Included Symptoms" while symptoms that were typically irrelevant were categorized as "Excluded Symptoms". Participants could exclude a typically included symptom by clicking the 'x' next to the symptom name, or include a typically excluded symptom by clicking on the '+' next to the symptom name.

Minor Usability Changes Because creating a chief complaint would now be accomplished through a separate interface, we realized that it would be frustrating to the user to create a lesson plan, select patients, and then realize that the chief complaint they had in mind was unavailable. We decided to move the chief complaint selection to the first screen (the lesson plan) in place of the required interview category selection. In every scenario created in the study so far, over all four prototype iterations, the user opted to include every interview category as required, so we decided that the SIDNIE system should always present all categories as options for selection.

Since both participants requested to be able to see the character in more detail, we added buttons labeled left, right, up, down, zoom in, and zoom out that would allow the participant to change their viewpoint in the character. Additionally, when the user opted to create a character, it was automatically selected as the character to use, instead of added to the library where the participant had to select it. Similarly, if the participant added in an option for the history of present

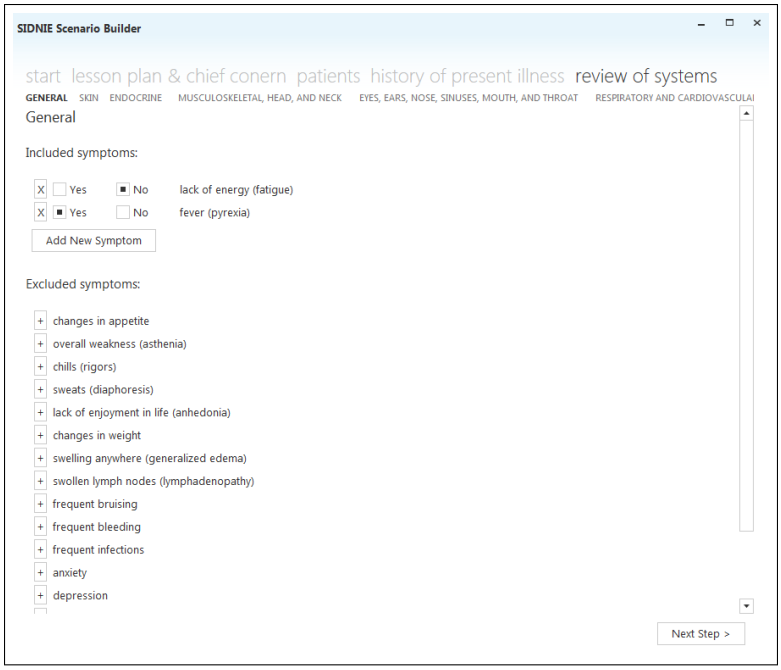


Figure 6.30: In the improved review of systems pages, symptoms are separated out into those typically relevant to the case ("Included Symptoms") and those typically irrelevant ("Excluded Symptoms").

illness or added in a symptom to the review of systems, it was automatically selected. Finally, for the electronic medical record screens for allergies, medications, and vaccinations, we added buttons and dialogue boxes that permitted the user to add an item instead of relying on the "blank line" cue. See Figure 6.31 for a screenshot of the improved electronic medical record screen for allergies.

6.6.7 Participant 3 Interaction

With the restriction of being unable to create a new chief complaint, the final participant created a scenario of a child with an earache because that was the only available chief complaint in the database. The participant filled out the lesson plan without any problems. When it came to the patient creation interface, the participant found the animation the patient was playing "a little distracting". Additionally, when she wanted to take a closer look at the patient she tried to click and drag within the window to change her perspective on the scene.

During the history of present illness, the participant commented that the list of words to describe pain quality was incomplete, and suggested that we add in a standardized set of words to

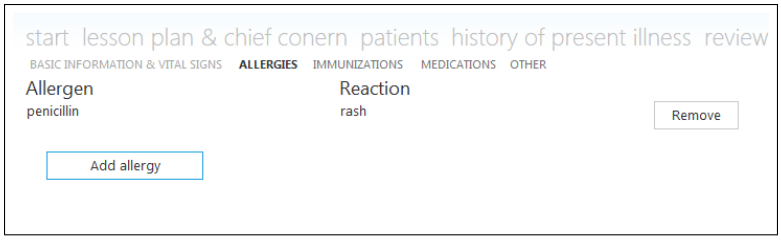


Figure 6.31: This is the electronic medical record's improved interface for the patient's allergies. Clicking on a the "Add Allergy" button launched a dialog box where the user could put in the allergen and reaction.

describe pain. In the review of systems, the patient was confused by the labels "Included Symptoms" and "Excluded Symptoms", since she thought that included meant that the symptoms were present, and excluded meant that the symptoms were absent. She suggested that the labels be changed to "symptoms typically pertinent to case" and "symptoms typically unimportant to case". She filled out the electronic medical record without any problems.

When it came to the question previewer, like the first participant, she expected to be able to navigate between question categories using the "next" button instead of by clicking through the tabs. She also did not understand that the questions were options for student selection within the scenario, pointing to a need to better describe the purpose of the questions in that interface. She noticed that some of the questions were not phrased how she would like them, and we explained to her that the questions were designed as a framework for her to then customize according to her learning objectives. She liked the idea of the automatically generated questions, saying that it was a great reduction in workload to "not have to think of all these questions" and match them up with the corresponding answers in the patient's case, but only to edit them according to her needs.

Overall the participant liked the new prototype, commenting that removing the previewer on the left hand side of the screen made it "easier to follow". She suggested several minor changes in wording of instructions. We explained to her the concept of outsourcing the database contribution to a different interface and possibly different users and she agreed that it was a good idea due to its streamlining of the scenario creation process.

6.7 Discussion

Besides the technical and usability errors that the participatory design process revealed, there were two crucial and unexpected paradigms that came to light. First, the nurse educators used the scenario builder tool in an unexpected way. Although nurse educators were able to use the "wizard" approach in that it provided a step-by-step template for all the information they needed to enter, they used it differently than a typical wizard since they wanted to be able to work on their patient scenarios in multiple sittings and occasionally wanted to move back to previously completed steps. This changed the entire design of the system and added significantly to its capabilities. This early observation would not have been possible within a laboratory setting since the task would have been much more constrained.

The second unexpected finding was that there are two separate pieces to scenario creation: creating a specific scenario, and contributing content back to the databases. Additionally, these two separate tasks could have two different target users. Only nurse educators can create scenarios, but many individuals are qualified to contribute content to the database. Separating the tasks streamlines the scenario creation process since the scenario creator can focus only on their desired scenario. The nurse educators did not help design an interface for content contribution back to the database, although their feedback with regards to the typical review of systems and creating and scoring questions will carry over into the design of the content contribution interface detailed in the following chapter. This idea was not unearthed until there was controlled observation and discussion of the prototype in action.

The final prototype in this study still had several usability difficulties, which we address in the following chapter detailing a usability study of the completed scenario builder tool.

6.8 Recommendations for Future Participatory Designs

Participatory design sessions are conducted in a variety of formats, ranging from focus groups to paper prototypes. Because participatory design involves all stakeholders, a participatory design session often includes both the creator of the product as well as its users collocated and working synchronously to design the product. This study differed some from this typical model in that two prototype feedback sessions occurred with the user completing the task on their own time, in their own setting of choice, and the creator receiving feedback afterwards. Those two sessions yielded

valuable insights that likely would not have been discovered in a controlled laboratory setting, or even in a collaborative setting, due to the freedom of task choices that participants experienced when using the software on their own time. Future product designers might consider at least one iteration of in situ use during the product design phase of the software lifecycle to capture user's habits and desires early in the implementation process.

Additionally, a theme that emerged during the design phase was the inverse relationship of the amount of time available for task completion and the user's skill level. Nurse educators, who have all the domain knowledge to create an accurate medical scenario, unfortunately do not have a lot of time to spare to do so. Nurse educators requested as much prefilled information as possible, yet did not want to take the time when creating the scenario to contribute content back to the database for others to use. In the end, nurse educators suggested that those with less skill and more time to spare (such as graduate or undergraduate assistants) could contribute content to the database, with nurse educators having the final say on their own scenarios' content. This created a new user group to design for. In the future, when faced with designing an interface for a complex, lengthy tasks, product designers might consider how adding another group of stakeholders or users could reduce workload for the most busy users while allowing for meaningful, skill-appropriate contribution from users who have more time on their hands. Especially with the advent of Amazon Mechanical Turk and other similar crowdsourcing systems, it might be that anonymous workers could complete some tasks at a low cost, freeing up skilled workers to tackle the more specialized or difficult tasks.

Chapter 7

Usability Evaluation of the Scenario Builder Tool

Although the scenario builder tool had been collaboratively designed with three nursing faculty members, we wanted to ensure that all usability concerns were resolved and that other nursing faculty not involved in the design would be able to use the system.

7.1 Changes to Scenario Wizard for Usability Study

Not all of the usability concerns of the participants were addressed by the end of the design process. Specifically, none of the changes the final participant suggested had been implemented and tested. We made the changes participant 3 suggested, and also smoothly integrated the SIDNIE system into the question preview interface to make the process clearer.

Participants suggested the removal of the nonessential "creation" tasks from the flow of the scenario builder tool. We implemented new interfaces for creating chief complaints and creating new scoring criteria. Although the character creation could have also been moved to the content contribution interface, all three participants in all three prototype interactions expressed that selecting or creating the character was their favorite part of their interaction with the system. Since it only takes approximately five minutes to create a character and seems to contribute significantly to the "fun" aspect of scenario creation, we chose to leave it within the wizard interface.

7.1.1 Searching For, Creating, and Customizing Characters

Due to button placement, in the previous study, there was often some confusion about how to proceed through creating and/or selecting a character. Since all other opportunities to add new content to the scenario builder tool occurred within the context of a modal dialogue box, we also transferred the process of creating a new character into a modal dialogue box. Inside the dialogue box, the "Cancel" and "OK" buttons were visible at all times, although the "OK" button was disabled until the character had been created. Clicking the "Cancel" button at any point in the process returned the user to the character library, while clicking the "OK" button at the end of the process added the created character to the library and selected that character for use in the scenario.

In the previous study, participants also expressed confusion about the buttons labeled for changing the style of the hair, eyebrows, and eyelashes when customizing a character. We replaced these buttons with a "thumbnail preview" of what the selected style would look like on the character, and placed buttons to the left and right of the thumbnail to give them the option to rotate through the available styles. Each style started out with an image that said "None", then the buttons allowed the user to loop through the available styles. See Figure 7.1 for a picture of the improved customization format. Additionally, users could zoom in or zoom out to view their character better, and could also rotate their character clockwise or counterclockwise to see the character's profile or back.

To increase the character realism in response to complaints during the SIDNIE study presented in Chapter 8 (that study was actually conducted before the study described in this chapter), We also updated the character generation software to make use of more realistic textures. The MakeHuman software provides eighteen high-fidelity skin textures for each combination of male/female, African/Caucasian/Asian, and young/middle-aged/old. we adapted the automatic character generation software to select the high fidelity skin texture that was most suitable for each generated character. we additionally adjusted the parameters of the software to allow for more facial differences and assymmetries from character to character to provide a greater variety of characters. Figure 7.2 shows the faces of three children generated from the updated software, while Figure 7.3 shows the faces of three adults.

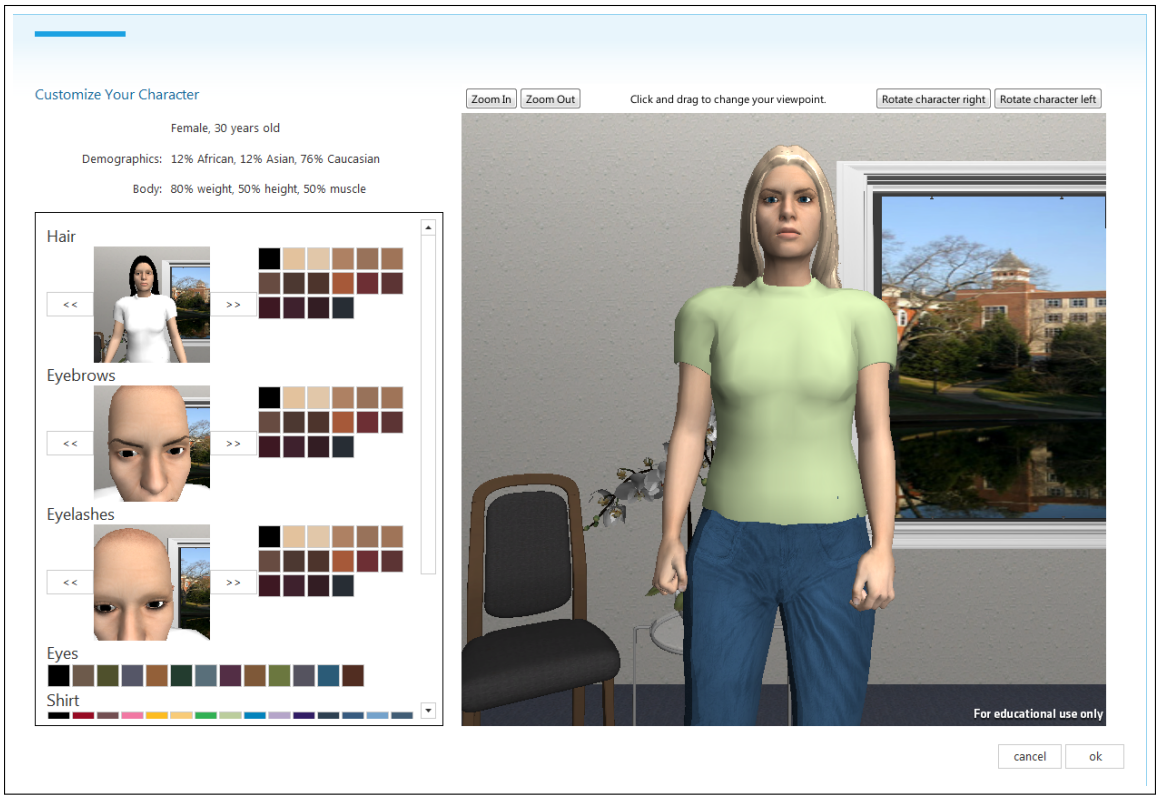


Figure 7.1: The updated customization interface shows thumbnail views of hair, eyebrows, and eyelashes with buttons to right and left to change their styles.

Figure 7.2: This figure shows three children created using the updated character generation software.

Figure 7.3: This figure shows three adults created using the updated character generation software.

7.1.2 Clarifying Question Previews through SIDNIE Integration

For two out of the three participants in the design study, the question previewer interface was confusing. Participants thought they should be removing all imperfect questions and did not understand that the questions were designed to act as options for students to select. To clarify this difference as well as to offer a more holistic preview option, we integrated a portion of the SIDNIE system into the question previewer interface.

When the previewer first is loaded, the user is prompted to position the pediatric patient by clicking and dragging to translate the patient up and down or side to side along the examination bench. The user can also rotate the patient using buttons labeled "Rotate Left" and "Rotate Right". After the child is positioned, the user can position the parent similarly. This allows the user some autonomy in deciding how to best stage the scene for the scenario. After both patients are positioned, the potential questions and answers appear below the characters. As in previous prototypes, the user can change the questions and answers however he or she likes. If the nurse clicks the "Preview this Question" button, then the virtual characters speak the answer to that question through text-to-speech with lip-syncing animations. See Figure 7.4 for the updated previewer layout.

7.1.3 Contributing a New Chief Complaint

The previous interface for creating a new chief complaint (Figure 6.28) required the user to review every symptom, and also did not give the user the option to mark each symptom as to whether it should be included or excluded in the review of systems, as the updated review of systems

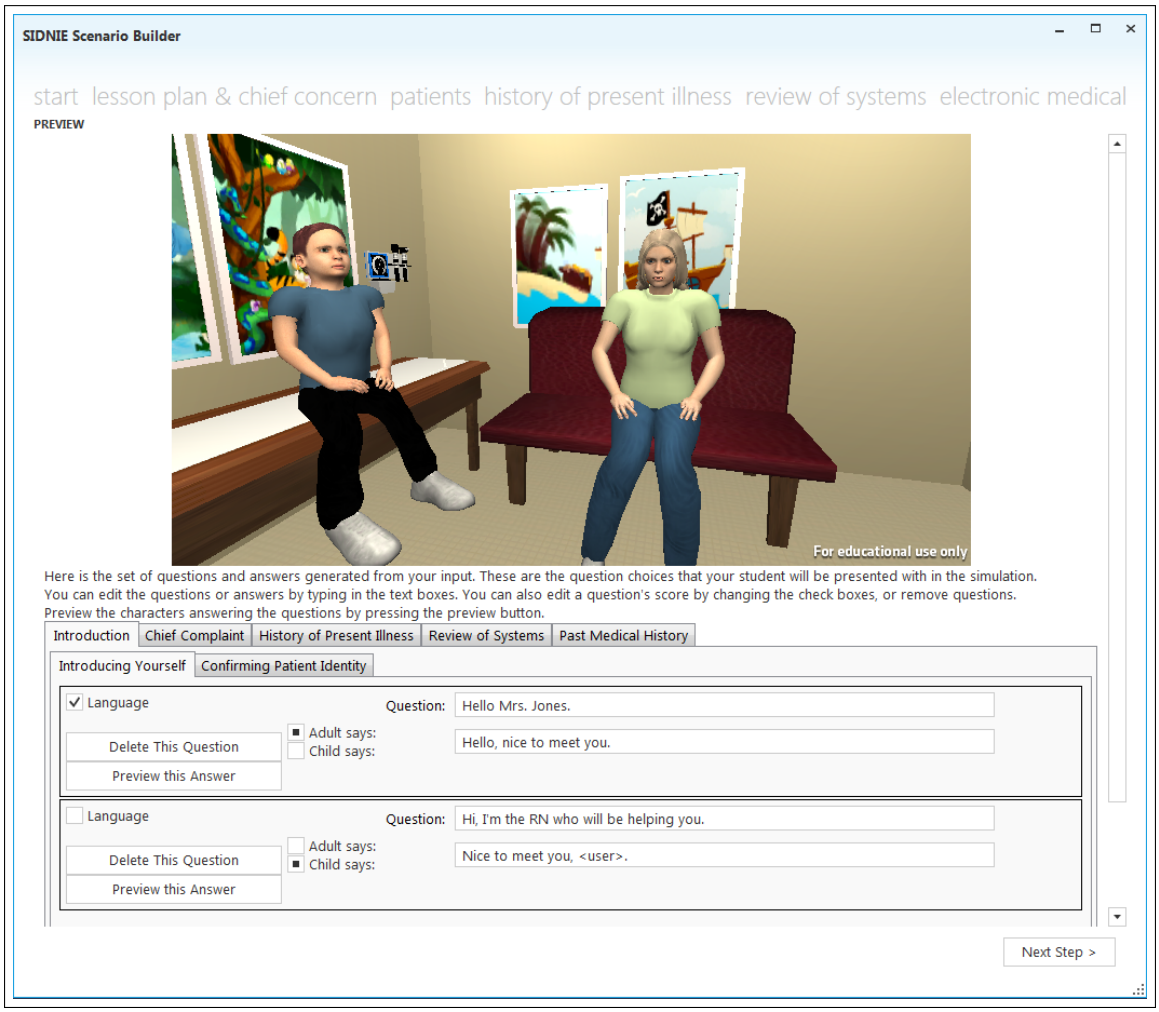


Figure 7.4: The updated previewer shows the selected patients in a doctor's office along with questions and answers. When the user clicks the button to preview a question, the patients respond with text-to-speech and lipsyncing.

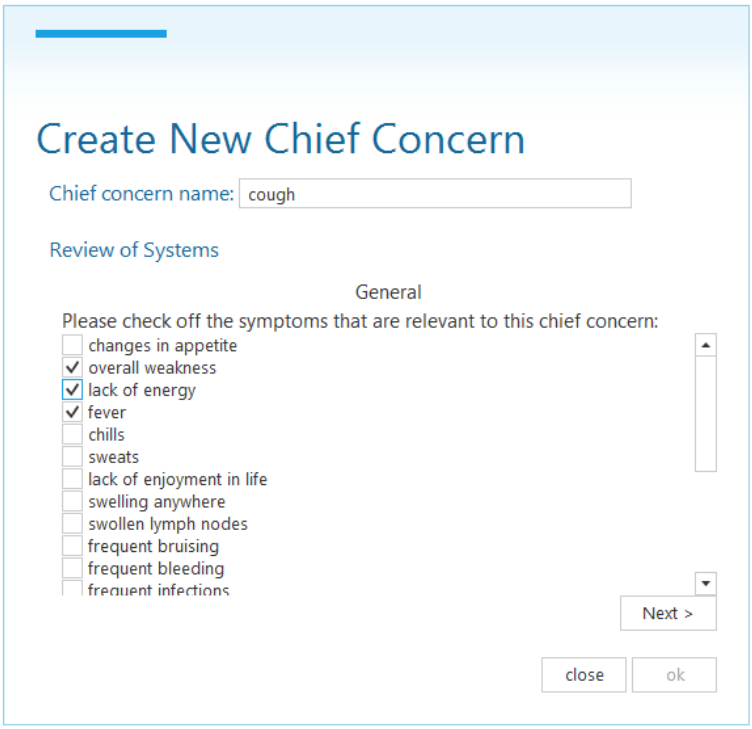


Figure 7.5: To create a new chief complaint, for each system, the user first checks off the symptoms they consider to be relevant to the chief complaint.

requires. We created a new interface for contributing a chief complaint. Once the user wrote the chief complaint's name, the first system's symptoms were displayed. The user then checks off each symptom he or she deemes relevant to the chief complaint (Figure 7.5)–these symptoms would be marked as included in the review of systems. Once the user clicks the "Next" button, each symptom marked as relevant was presented one by one to the user, where the user had to mark each symptom as typically present or typically absent (Figure 7.6).

After completing all the relevant symptoms for that system, the process repeats for every remaining system. Finally, the user can review all their selections and edit them if necessary. When the user clicks "OK", the dialogue box closes and the chief complaint is added to the database of available chief complaints.

7.1.4 Contributing a New Scored Criteria

For contributing a new scored criteria, first the user enters the criteria's name and definition. Then, for each category in the interview, the user completes the following three tasks. See Figure

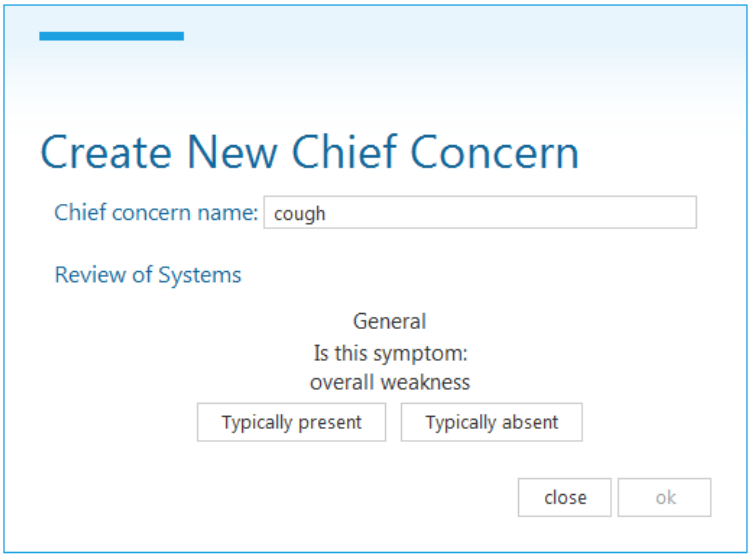


Figure 7.6: After selecting the relevant symptoms, the user selects whether each relevant symptom is typically present or typically absent for their chief complaint.

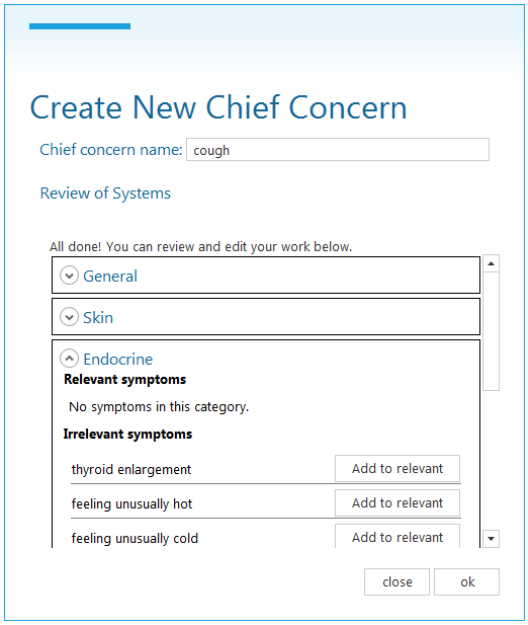


Figure 7.7: After selecting the relevant symptoms, the user selects whether each relevant symptom is typically present or typically absent for their chief complaint.

7.8 for a screenshot of the interface after the name and definition are entered.

- *Check off the existing questions in the database that meet the new criteria.* Each existing question was listed with a checkbox to its left.

- *Write a question that meets the new criteria, along with all the other criteria in the database.* This step ensures that for each category in the interview, a "perfect" question can be selected by the user, regardless of the combination of scoring criteria selected for that scenario. In addition to a text box where the user can type question text, there is a drop down menu labeled "Insert template..." that contains relevant placeholder text for each interview category that can be inserted inline with typed text. For example, in the category "Introducing Yourself", there are templates offered for the child's name and the parent's name, presented as <child> and <adult>. When that question is selected for a scenario, the templates would be filled in with the appropriate text given in the scenario.

- *Write a question that does not meet the new criteria, and score it for all remaining criteria.* This step ensures that for each category in the interview, it is possible to select a question that does not meet the new criteria. Scoring the new question for all other existing criteria makes the question usable for other scenarios that do not include the new scoring criteria, leading to a greater variety of questions for every possible criteria combination.

After completing every interview category, the user clicks the "OK" button and the criteria and written questions are contributed to the database for use in future scenarios.

7.1.5 Additional Concerns Addressed

In the section where the user selected the quality of the pain, we added descriptive words for pain from the McGill Pain Questionnaire [Melzack and Katz, 2007], a medically-accepted measure for pain. We changed the labels for symptoms that should be included or excluded in the review of systems to "Symptoms to ask about in relation to (the chief complaint's name)" and "Symptoms to ignore". We added input validation for each page so that a user could not proceed without completing each question, and added borders around questionnaire sections to increase readability.

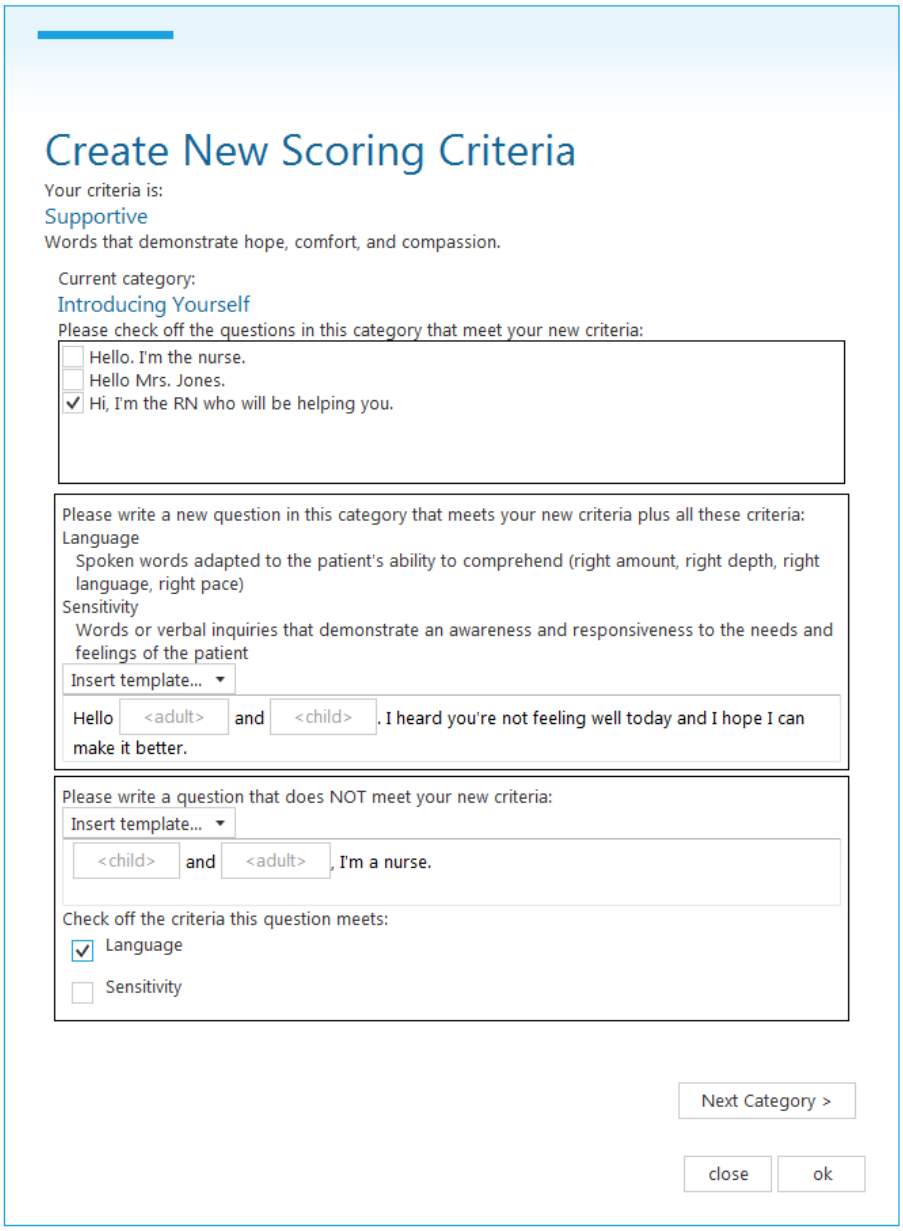


Figure 7.8: When creating a new criteria, for each interview category the user is required to score all the existing questions as well as to write two new questions to contribute to the database.

7.2 Experimental Procedure

Six Clemson University School of Nursing faculty and one Clemson University upper level nursing student were recruited to determine the usability of the scenario builder system, since usability research indicates that seven participants are sufficient to find the majority of usability problems in a small project [Faulkner, 2003, Nielsen and Landauer, 1993]. Two participants in this study also were participants in the study for the design of the scenario builder tool. The remaining four faculty members had never interacted with the scenario builder tool before. The student participant was recruited because in the previous design study, a participant suggested that an undergraduate or other assistant might be able to do some of the more time consuming tasks involved in generating a scenario.

Each participant signed an informed consent which detailed their involvement in the research process. Once informed consent was received (Appendix B), the participant answered questions about their nursing experiences and experiences with children (Appendix C).

I then showed each participant who had not interacted with SIDNIE before a video of a virtual patient system to ensure they knew what a virtual patient was and understood the student's interaction with the SIDNIE system.

Each faculty participant completed the following three tasks:

1. Create a scenario for a specific case study. We provided a paper copy of a case study with a specified chief complaint and scoring criteria (Appendix F). The scoring criteria and chief complaint for this case study were already present in the scenario builder tool library, so they did not have to add a scoring criteria or chief complaint to get started. We gave the participant as much time as they would like to read the case study, then asked them to create a scenario for that case study, using their medical knowledge to fill in any information that was not specified in the case study.

2. Create a new chief complaint. We provided a paper copy of the typical review of systems for the chief complaint, including which symptoms were pertinent to the case and which were not, but instructed the participant that they could fill it out using the sheet of paper or using their own nursing knowledge.

3. Create a new scoring criteria and corresponding questions. We provided a paper copy of the

criteria and its description (the criteria *supportive* was selected from the validated rubric shown in Appendix D). Because this process can be time-consuming and the study was limited to one hour, each participant only completed the scoring and writing for two categories.

The student participant completed an alternate first task, with similar second and third tasks:

1. Complete the "Previewer" portion only of an otherwise-created scenario. Since this task requires going through question by question to modify answers to suit the patient demographics (while requiring little medical knowledge since question and answer templates are already filled according to the input to the previous steps in the scenario builder), it is a candidate task for someone other than a nurse educator to complete. We provided the student with a copy of the same case study (Appendix F) as used in the faculty task.

2. Create a new chief complaint. As in the faculty task, we provided a paper copy of the typical review of systems for the chief complaint, including which symptoms were pertinent to the case and which were not. We instructed the student to complete the task based on the information on the sheet, since in actual usage a nurse educator could provide an existing typical review of systems from a textbook or other source so that the student would not have to rely on his or her limited medical knowledge. Participants in the design study suggested that this task could be completed by a student.

3. Create a new scoring criteria and corresponding questions. We provided a paper copy of the criteria and its description (the criteria *supportive* was selected from the validated rubric shown in Appendix D). Participants in the design study suggested that this task could also be completed by a student. Instead of completing two interview categories, the participant completed three interview categories.

We did not do a demonstration or tutorial of the scenario builder system before interacted with it because we wanted to be able to informally gauge its learnability as well as see what stumbling points emerged naturally if there were no instructions. We observed and videotaped each participant as they interacted with the system. The system automatically recorded the time taken for each step of the scenario creation process and took a screenshot of the final state of each step as well. We answered any questions the participants had during their interaction with the system, and if the

participant seemed to be confused any time during their interaction, we asked them to describe their thought process and then helped them move forward, while making note of any usability problems that arose or any need for instruction during the task.

Between each task, we administered a short post-task interview (Appendix G) to gather the participant's impressions on the system for completing that task in particular. After the participant completed all three tasks, we administered the System Usability Scale [Brooke, 1996](Appendix I) and a post-interview to gather information on the specifics of how participating nursing faculty perceived the system (Appendix H).

My hypotheses for this experiment were:

1. The participants will find the system easy to use.

2. Faculty participants will be able to make a scenario in less than 30 minutes.

7.3 Results

Seven participants took part in the experiment–one student, and six faculty members. Participants varied widely in every demographic measure, with some participants reporting little experience in any of the measured categories, and some reporting high experience in all categories. Two participants previously participated in the design of the scenario builder tool, while the remaining participants had never used the scenario builder tool before. See Table 7.1 for the demographic information of each participant.

Table 7.1: Demographic information for the seven participants in the usability study. In the "Role" row, F stands for faculty member, while S stands for student.

Participant Number	1	2	3	4	5	6	7
Role	F	F	F	F	F	F	S
Helped with design?	No	No	Yes	Yes	No	No	No
Teaching undergraduate courses	5	4	3	5	1	5	1
Pediatric patients	3	4	3	3	1	5	2
Computers	5	4	5	5	1	4	4
Virtual reality	2	2	5	4	1	2	3
Developing nursing simulations	5	2	5	4	1	2	2

7.3.1 Task Completion Time

Out of the three tasks, faculty participants took the longest to create the scenario. The system automatically recorded the amount of time spent in each scenario creation step. Due to a technical error the time was only recorded the first time that a faculty member visited a scenario step, so if he or she chose to return to a step the time was not recorded. However, only two of the six participants returned to previous steps, and from observation, when a faculty member returned to a previous step, it was a very brief interaction either to correct a mistake or to review what he or she had put on a previous step. Table 7.2 shows the time faculty members spent in each step of the scenario creation process, along with means and standard deviations. The final "Preview" step is excluded because due to usability problems, two participants skipped the step altogether, and only two other participants looked at more than two questions in the previewer instead of looking through all the questions in the interview, as intended. On average, participants completed the scenario in approximately 23 minutes.

Table 7.2: This table shows the amount of time faculty participants took to complete each step in the scenario creation process in minutes and seconds.

Participant Number	1	2	3	4	5	6	Mean	SD
Lesson Plan	02:19	02:27	01:35	01:13	01:45	03:46	02:11	00:54
Patients	08:35	09:08	03:25	04:40	07:05	10:15	07:11	02:40
History of Present Illness	05:43	05:27	04:49	07:05	05:46	05:51	05:47	00:44
Review of Systems	01:56	07:25	06:30	03:50	03:01	12:53	05:56	04:00
Electronic Medical Record	00:27	01:52	02:32	04:01	00:23	02:58	02:02	01:26
Total	19:00	26:19	18:51	20:49	18:00	35:43	23:07	06:52

The remaining two tasks took less time to complete. Participants could create a chief complaint in approximately four minutes. For the scoring criteria, due to time constraints we only required each participant to complete two of the eleven possible interview categories. Participants could complete two categories in approximately six minutes, leading to a projected completion time of all categories of approximately 32 minutes. See Table 7.3 for the recorded completion times for each faculty participant.

The student participant completed an alternate set of three tasks. The first task, working through the scenario previewer, took the student 25 minutes and 6 seconds. The student checked each question and answer that was generated by the scenario builder tool. The second task, creating a chief complaint, took the student 5 minutes and 32 seconds. The final task, completing three

Table 7.3: Completion times for the final two tasks for faculty members. The third column is a projected completion time for the task of creating a new criteria, since the task was only completed in part in the usability study due to time constraints.

Participant Number	1	2	3	4	5	6	Mean	SD
Create Chief Complaint	02:11	03:44	05:18	03:10	01:47	05:22	03:35	01:31
Create Scoring Criteria (Two Categories)	06:36	03:42	07:19	06:29	05:23	05:20	05:48	01:17
Projected Time for All Scoring Criteria	36:18	20:21	40:15	35:39	29:37	29:20	31:55	07:03

interview categories for creating a new scoring criteria, took seven minutes and four seconds, for a projected task completion time of 25 minutes and 55 seconds.

7.3.2 Task Difficulty Ratings

We asked the participants to rate the difficulty of their tasks in two ways. First, between each task we asked them to rank the difficulty of the task they just completed on a scale of 0 (very easy) to 5 (very difficult). Table 7.4 shows each user's task ranking. Mean task rankings all fell below 3 (medium difficulty). The student participant ranked the difficulty of all three of his tasks as 1 (least difficult).

Table 7.4: Difficulty rankings for each task completed by faculty participants, where 1 is the least difficult, and 5 is the most difficult.

	Scenario	**Chief Complaint**	**Scoring Criteria**
Participant 1	2	1	2
Participant 2	2	1	2
Participant 3	2	1	2
Participant 4	2	2	3
Participant 5	1	1	1
Participant 6	4	4	3
Mean	2.17	1.67	2.17
SD	0.98	1.21	0.75

Second, in the post-task interview, we asked participants to name the most difficult task and the least difficult task, providing a comparative ranking. Four out of the six faculty participants found the scenario creation task the easiest. Similarly, four out of six faculty participants found the scoring criteria task the most difficult. See Table 7.5 for the overall summary rankings. The student participant found creating a new chief complaint least difficult and creating a new scoring criteria most difficult.

Table 7.5: Faculty participants reported which task was the most and least difficult.

	Most difficult	Least difficult
Scenario	2	4
Chief Complaint	0	1
Scoring Criteria	4	1

7.3.3 System Usability

Each participant filled out the System Usability Scale. We instructed participants to consider all three tasks together when completing the scale. Average scores for each question were at 3.85 or above out of 5 (with 5 indicating the highest usability), with all scores reversed to match question phrasing. The lowest scoring question was "I think I would need the support of a technical person to be able to use this system" (reversed mean=3.86, sd=0.90), while the highest scoring question was "I found the system very cumbersome to use" (reversed mean = 4.43, sd=0.79). Table 7.6 shows the scores each participant gave each question along with means and standard deviations. When converted to percentage scores, the mean SUS score was 83.43% (sd=10.05), indicating overall good usability.

Table 7.6: Responses for the system usability scale by participant.

Participant Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Mean**	**SD**
I think that I would like to use this system frequently.	5	4	5	5	2	4	4	4.14	1.07
I found the system unnecessarily complex.	5	3	4	5	5	3	4	4.14	0.90
I thought the system was easy to use.	5	4	5	4	5	3	4	4.29	0.76
I think that We would need the support of a technical person to be able to use this system.	4	3	4	3	5	5	3	3.86	0.90
I found the various functions in the system were well integrated.	5	4	4	4	5	3	3	4.00	0.82
I thought there was too much inconsistency in this system.	4	4	5	5	5	3	4	4.29	0.76
I would imagine that most people would learn to use this system very quickly.	5	5	3	4	5	4	4	4.29	0.76
I found the system very cumbersome to use.	5	3	5	5	5	4	4	4.43	0.79
I felt very confident using the system.	5	3	5	5	5	3	3	4.14	1.07
I needed to learn a lot of things before I could get going with this system.	5	4	4	3	5	5	3	4.14	0.90
Percentage	96	74	88	86	94	74	72	83.43	10.05

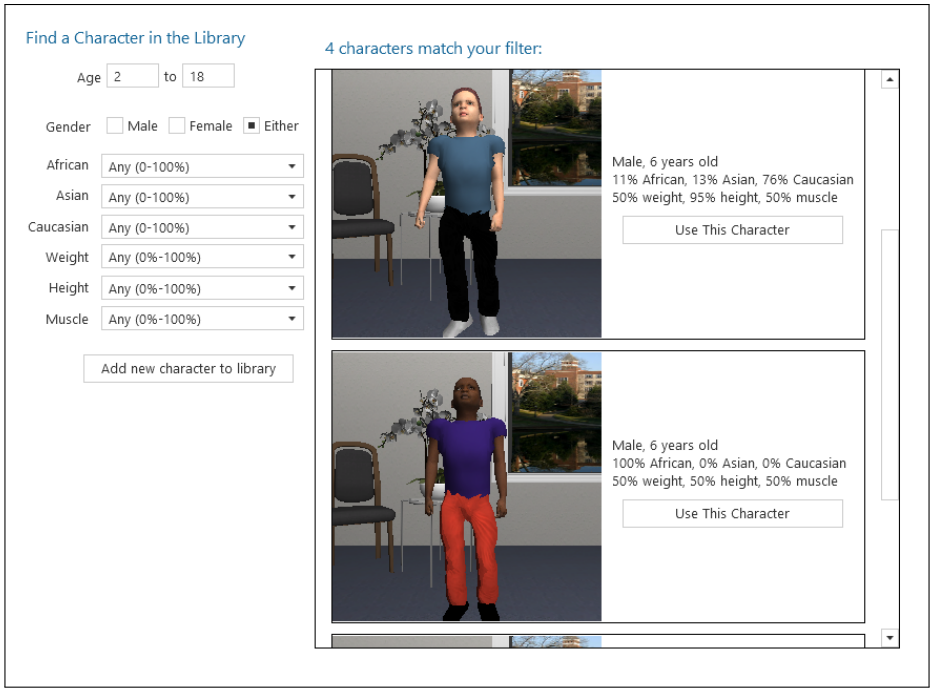


Figure 7.9: The user can select a patient from the library or choose to create their own character.

7.3.4 Usability Observations

Informally, participants gave feedback throughout their interaction with the system as they asked questions and explained their reasoning as they encountered difficulties. Participants tended to have the same usability problems throughout the task interaction.

7.3.4.1 Character Selection and Creation

Every participant had some difficulty in selecting or creating a character. When the participant first enters the interface to select a character, he or she sees a patient library with filters to the left (Figure 7.9). As filters are modified, the contents of the box on the right side of the screen are immediately modified to show only patients that match that filter. From this point, there were two distinct usability flaws. First, at least three out of the six faculty participants did not understand that the library was immediately filtered to meet their constraints, but instead expected there to be a "Search" button that would make their filters take effect. Consequently, once the user filled out the filters, they were uncertain what to do next. Usability could possibly be improved by adding a search button and disabling the immediate filtering.

Second, at least two of the six faculty participants filled out the filters, clicked the button to create a new character, then became frustrated that the information they filled out in the filter did not carry over to the dialogue box where a new character is created. The dialogue is instead populated with average values (Figure 6.26). We had initially designed it this way because most of the filters are presented as a range, while character creation requires specific values (instead of a range) to create the character. A possible solution would be to randomly select values within the specified ranges to automatically populate the dialogue for creating a patient, then ask the user for confirmation before creating the character.

We observed that all participants, when creating a character, instead of filling out the dialogue in order of its fields (Figure 6.26), filled out the gender, ethnicity, then age, and finally the body characteristics. This suggests that the controls should possibly be rearranged on the page into that ordering. Additionally, three participants reported confusion about the "muscle" classifier. This is intended as a general measure of muscle tone. One participant interpreted it as actual body composition percentage, noting that "not very many people have 50 percent muscle", indicating that it would be physically impossible to get to higher percentages. Two participants said that they would not think about a patient's muscle percentage in this context. This feedback suggests that it might improve usability to exclude the ability to edit muscle percentage, and instead to make it correlated with weight, where patients with low weight would automatically also show higher muscle tone and patients with higher weight show lower muscle tone.

When customizing the character, the primary difficulty was in the interface for selecting hair, eyebrows, and eyelashes (Figure 7.1). Every participant clicked the hair color before selecting a hair style, leading to confusion since there was no hair visible. To improve usability, the hair colors could be hidden until a hair style is selected. Four out of six participants clicked the button to add a hairstyle once, moving to the first hairstyle, and then stopped. When asked whether they realize they could move beyond the first hairstyle, the participants said they did not understand that, with one participant saying that she thought the button "put the hair on the patient" and if clicked again might "take it off". This suggests that the buttons should be labeled differently.

Each participant made a character with one race highly predominant, although one participant noted that she liked that you could make a lot of different cultural situations by mixing the races. A participant commented that she would also like to see different hairstyles, particularly "dreads or twisties". Additionally, participants enjoyed seeing the patients being animated during

the customization step, with one participant saying "I like the way he moves!" Only one participant chose to zoom in to view the patient closer, and only one participant rotated the patient.

7.3.4.2 History of Present Illness

Overall, this section seemed to be fairly straightforward for most participants. The only difficulty encountered was that two participants were unsure how to fill out the section on location, quality, and severity of pain because in their opinion the case study did not indicate that the patient was experiencing any pain. To improve usability, it could be possible to ask whether the patient was experiencing any pain, then skip that portion if there was no pain reported. In the same section, at least three participants clicked on the faces pain scale image to set its level instead of typing the numerical pain score in the text box. The pain scale could be easily made into a clickable control to aid usability.

7.3.4.3 Review of Systems

Out of all the interview sections, this portion received the most diverse feedback. Four out of six faculty participants reported some level of difficulty in understanding the review of systems and what it meant. Three of the six faculty stopped on the first body system ("General") and immediately started adding new symptoms for other body systems (such as cough, runny nose, or stuffy nose, which would appear in the "Head, Ears, Eyes, Nose, and Throat" system). Two participants were confused about the included and excluded symptoms, particularly in combination with the yes/no option for whether the symptom was actually present or absent. One participant figured out the interface by clicking buttons to decide what they did, while at least one participant missed the visual cue that happened when a symptom was moved from one section to the other.

7.3.4.4 Electronic Medical Record

There were no obvious problems with the electronic medical record. Three out of six faculty participants left it unmodified from default. One participant added a medication, one participant added an image, and one participant modified the vital signs to better fit the case.

7.3.4.5 Previewer

Participants enjoyed placing their characters in the doctor's office and seeing them animated in the room. However, only one participant spent a significant amount of time looking over questions in the previewer, visiting each category. It seemed that by the time the participant reached this page they were ready for the task to be completed and were rushing through to finish. Two participants skipped the previewing step altogether, while the remaining two participants only looked at one or two questions before ending their interaction. The faculty participants who did spend significant time with the previewer edited questions and answers for grammar, and often tailored the answers to better fit the patient's demographic. Nearly all participants did not initially see the tabs for interview subcategories. Participants seldom used the button to preview the question by having the characters speak the answer, but when they did, they commented on the poor quality of the text to speech voices.

The student participant spent an extended time with the previewer interface. His first feedback was that he needed better instructions to understand the task–in particular, the question scoring. The student looked at every question and answer. The majority of time spent was in editing the answers to better fit the patient's demographic. Occasionally the student changed the scoring for a question or deleted a question that was irrelevant or redundant.

7.3.4.6 Creating a New Chief Complaint

There were no outstanding usability concerns with the interface for creating a new chief complaint. Five of the six faculty participants chose to fill it out from their own medical knowledge, while one faculty participant as well as the student participant filled it out by referencing the provided chart.

7.3.4.7 Creating a New Scoring Criteria

Of all the tasks, this one had the most usability problems, likely due in part to its overall complexity and lack of presence in the design process. All participants had some degree of difficulty with understanding the task and the interface.

Of all seven participants, only one participant (the student) used the provided templating ability, but inconsistently–the student used templates for two questions, then typed scenario-specific

information in the third question. The remaining six participants wrote their questions either carefully avoiding any information specific to the scenario (questions such as "Hi, I'm the nurse who will be taking care of you today, what is your name and birthdate?"), using the names they had chosen for their characters in the scenario they created, or using the default information used to fill in the question templates in the previewer interface for the questions that needed to be scored for the new criteria (Figure 7.8). Participants also got confused about which instructions corresponded to which questions, often unintentionally skipping sections.

This task also seemed to cause a high mental workload. It often took a couple minutes for a participant to come up with a question that met all the criteria. Writing a question that did not meet the criteria seemed less difficult, although scoring that question also seemed to be a challenge.

7.3.5 Subjective Feedback

Participants also answered questions about usability in a post-task interview. Most participants suggested the use of a tutorial for one or more parts of the interaction in response to at least one question in their interview. Overall, feedback was positive, with five out of the seven participants mentioning that the process was user friendly. When asked what participants liked the best about their experience, two participants mentioned the avatars, four participants said they thought it was quick, easy to use, or thorough, and one participant said they thought it was just a good idea to be able to "create an interactive dialogue and practice without an actual human being".

When asked what they liked the least and what they would like to change, three participants suggested adding a tutorial for one or more tasks. Four participants cited usability problems they had encountered during their interaction. One participant suggested a different ordering for the interview, as a SOAP (Subjective, Objective, Assessment, Plan) note [EMRSoap, 2014] instead of the Medicare standard we based the structure on [of Health et al., 2014]. Two participants complained about the text to speech voices, and one said that making the new scoring criteria was time consuming. There were also suggestions for more content such as different hairstyles and more prexisting characters or chief complaints.

7.4 Discussion

My first hypothesis, that *the participants would find the system easy to use*, was at least partially confirmed. While each participant encountered usability problems, the SUS questionnaire results as well as their overall positive feedback during the final interview indicates that the system was usable overall. In addition, the feedback given during this study could be used to greatly improve usability.

While there were easily correctable usability problems in nearly every task and subtask, the majority of the significant usability problems fell in three sections: the review of systems, previewing the scenario, and the creation of a new scoring criteria. In the Review of Systems, participants tended to misunderstand the labeling, adding symptoms to body systems where they did not belong, and being confused about the difference between ignoring a symptom, asking about a symptom, and marking whether the symptom was present or absent. Conversely, there were no reported usability problems in the dialogue for creating a new chief complaint, which was an abstracted version of the exact same task, since for a new chief complaint each symptom had to be classified as irrelevant, relevant and typically present, or relevant and typically absent. Having preset chief complaints in the library was originally designed to save time during the scenario creation process, since users could rely on the previous work of the chief complaint creator and be presented with a default review of systems for their chief complaint for further customization. However, this usability study shows that participants spent 356 seconds on average completing the review of systems, when they only spent 215 seconds creating a new chief complaint. This suggests that in a future iteration, removing the chief complaint "library" and requiring the user to fill out a review of systems from scratch using a similar interface may actually be a time-saving measure.

Faculty participants seemed to be in a hurry to finish the task by the time they got to the previewer interface, with only one faculty participant taking the time to look at and modify more than two questions. Given the corresponding case study, the student was able to successfully use the interface to tailor questions and answers to the scenario and remove unnecessary questions after receiving some instruction about the task. This suggests that the task of previewing the scenario might be better split apart from the rest of the scenario generation task and saved for later review by the nurse educator or another individual.

Finally, the task of creating a new scoring criteria was a significant challenge for every user.

Besides the interface difficulties that are easily corrected, the task of creating questions seemed to be difficult overall. In all three tasks, participants could use provided reference materials. However, for this task, the only reference provided was the criteria name and definition. This increased the task complexity since the user had to both figure out a new interface and come up with questions that met or did not meet the new criteria. It is possible that the task would be easier given more time, or given an "offline" way of completing the task before introducing the complexity of the interface. For example, the system could provide a spreadsheet of all existing questions to score for a new criteria, and prompts for coming up with the new questions. Then a user could complete the work at his or her own pace and enter the data later. Additional usability studies focusing on this task could also reveal better ways to reduce cognitive load and support task completion.

For both the previewing and scoring criteria tasks, faculty participants expressed concern at the amount of time either task would take. Since the student participant was capable of completing each of these tasks, it is possible to reduce that workload by outsourcing those tasks to another individual. One participant said, "Time is a big factor when it comes to our work...creating the avatar is fun; the other stuff is time consuming...let an undergraduate do the nitty gritty and then let the faculty member sign off at the end," suggesting that there could be a step after student tasks where it the scenario or criteria is finalized and endorsed by the nurse educator or faculty member.

The second hypothesis was that *faculty participants would be able to make a scenario in less than 30 minutes.* Excluding the preview stage, five of six faculty participants were able to complete the scenario in less than a half-hour (average time was 23 minutes, 7 seconds, with a standard deviation of 6 minutes, 52 seconds). If the previewer task is outsourced to another individual, then this hypothesis holds true for a scenario where the chief complaint and scoring criteria are already present in the library. This task completion time ignores the time spent finding a case study and becoming familiar with it; however, this cost is necessary whether using the scenario building tool or using a different simulation method such as a standardized patient or roleplaying. Keeping the interaction time under 30 minutes also can potentially reduce software and task complexity, since generally an individual can dedicate 30 minutes to a task without having to stop in the middle. This means that saving and loading scenarios in the middle of the "wizard" style steps is likely unnecessary.

Using the average for all task completion times, to create a new chief complaint, add a scoring criteria, create a scenario, and tailor the questions using the previewer, the total task completion

time would be 1 hour, 35 minutes, and 19 seconds, with approximately 1 hour of that time being able to be completed by someone other than a nurse educator. This is a dramatic improvement over the nine months it took to create one scenario in our experience.

In addition to usability, it is important to determine whether the scenarios created by the tool are actually useful in a learning context. In the following chapter, scenarios created in less than a half-hour using the scenario builder tool were imported into the SIDNIE system to determine whether they yielded positive learning outcomes.

7.5 Recommendations for Designing Content Creation Interfaces

Both the participatory design sessions and the study on the scenario builder's usability brought to light key concepts to consider in designing content creation interfaces. One factor to remember is the amount of mental workload and preparation that is necessary to create new content. In this case, the participatory design study showed that nurse educators chose to rely on already-familiar cases–one participant by looking up a case study online, and one participant by choosing a scenario that he was covering in class that week. In the usability study, the most difficult task for most participants was the task for creating a new scoring criteria. In the first two tasks, participants could rely on their own knowledge and on the reference material provided to complete the majority of the task. In the final task, however, it took participants quite some time to come up with questions on their own, even though they well-understood the criteria they should be scored by. In all three tasks, it seemed that the difficulty of generating new content was often conflated with the challenge of using a new interface, leading to confusion or frustration. In the future, it seems that the "offline" workload of creating content should be measured and considered as a part of the creation task, and that providing supplemental, offline resources (for example, worksheets or instructions) could potentially speed along the content creation process and reduce difficulty.

Another observation coming from both studies is that users often do not read written instructions. Throughout the multiple iterations of this software, we have rewritten field labels multiple times to improve clarity, and have added instructions in many places to disambiguate what the user should do. However, the response is nearly invariable: users start clicking buttons and observing the results to figure out how to use the interface, rarely reading the instructions. This leads to many user

errors and confusion if their action does not yield the expected results. Typical usability wisdom teaches to make changes obvious and to provide affordances to the user to be able to undo or cancel actions. Unfortunately, this helps little if there is an overall conceptual misunderstanding of the task. For example, in the previewing task in this study, several participants deleted questions that they thought were not good questions, while the point of the system was to provide questions that both met and did not meet the scoring criteria. Similarly, in previous iterations, participants filled out the chief complaint as it pertained directly to their specific case study instead of in a generalized sense. For both these tasks, there were clear written instructions and affordances for correcting mistakes, but participants rarely read the instructions or realized they were misunderstanding the task until it was pointed out to them verbally. It is possible that adding a non-interactive tutorial before the task begins could correct some misconceptions since participants would not have the option of experimentally figuring out the interface, but instead had to listen to instructions. In any case, future designers would do well to anticipate "experimentation" within their interfaces and should expect that very little instruction text will be heeded.

Finally, it is important to consider how the time needed to complete a task changes the user's expectations and requirements for the software. In the initial prototypes, users strongly suggested the ability to save and load scenarios as they worked on them, indicating that they could not complete the task in one sitting. Although we did not measure the amount of time users spent with the initial prototype, in this final usability study, participants completed the entire scenario generation task in less than a half-hour, which is a reasonable time to be completed in one sitting. None of the participants in this study (including those who participated the design study) requested the ability to save and load their scenarios for future work, and few participants were even concerned about returning to previous steps to check their work or reference what they had already put in in other fields. However, when reaching the previewer step, participants often hurried through it or skipped it altogether, suggesting that after about a half-hour they wanted to be able to finish the task. Additionally, no participants returned to previous steps after reaching the previewer. This suggests that the previewer task should be considered altogether separate so that it can be delayed to another time or outsourced to another user. Reducing the time taken to create a scenario led to a different problem and solution than originally suggested. As usability concerns are addressed and task completion time is reduced, interface designers should revisit original requirements to determine whether they are still relevant, and elicit new requirements as new patterns emerge.

Chapter 8

Evaluation of Learning Scaffolding

The goal of this experiment was to assess whether SIDNIE's scaffolding method is effective for learning verbal interviewing skills as measured by a validated rubric [Diers, 2008] (Appendix D), as well as increasing self-efficacy. We tested the working hypothesis through the use of:

1. validated methods of measuring self-efficacy [Chen et al., 2001] (Appendix L)
2. tests for learning retention (Appendix M)
3. validated instruments to measure verbal interviewing skills [Diers, 2008] (Appendix D).

A secondary goal was to evaluate whether basic scenarios made with the scenario builder tool could be effective for learning. We created each of the three scenarios used in this experiment using the scenario builder tool. Each scenario was based on a pediatric case study, and I spent approximately 30 minutes creating each scenario. We created new characters with suitable demographics for each scenario, and filled out the history of present illness and review of systems according to the case study reports. For the automatically generated questions and answers, we corrected any grammar errors in the automatically generated questions and answers, and in some cases tailored the answers to better fit the pediatric patient's age.

In previous studies, the virtual patient's animations were reported to be distracting from the case content [Bloodworth et al., 2011], and unnecessary to achieve learning outcomes [Pence et al., 2013]. To remove any confound that would be caused by animations that had not been empirically validated, the only animations displayed were breathing and mouth movement in synchronization with the text to speech.

My hypotheses were:

1. Participants who were in the scaffolded condition would show better learning outcomes than those in the non-scaffolded condition immediately after participation.

2. Participants who were in the scaffolded condition would show better learning retention two weeks after participation than those in the non-scaffolded condition.

3. Participants who were in the scaffolded condition would show a greater increase in self efficacy than those in the non-scaffolded condition.

8.1 Experimental Procedure

We asked each participant to sign an informed consent which detailed their involvement in the research process (Appendix J). Once we received informed consent, each participant completed a demographic questionnaire (Appendix K) to record the participants' relevant background experiences, such as experience with human patients, children, and virtual reality, since these factors may have affected the participant's performance.

Next, the participant completed the Generalized Self Efficacy Scale [Chen et al., 2001] (Appendix L) to evaluate their current level of self-efficacy. Then, the participant completed a pre-training questionnaire to gauge their current learning level (Appendix M).

Participants were randomly assigned to one of two groups: scaffolded or unscaffolded. Each participant interacted with three separate scenarios. In our previous work, we found that using the same scenario for multiple scaffolding levels created a confound for measuring learning outcomes. By the fourth exposure to the scenario, the students remembered which questions in the scenario were correct [Dukes et al., 2013]. To address this confound, we used a different scenario for each of four interactions (while using the same scoring criteria for each interaction). The scaffolded group will interact with SIDNIE using three scenarios with decreasing feedback and guidance over each scenario. The unscaffolded group interacted with SIDNIE using the same three scenarios, yet receiving only summative feedback at the end of each scenario.

Once the scenario interaction is complete, the participants completed the same questionnaire (now called the post-training questionnaire) they completed before they interacted with the system to gauge his/her learning level after interaction with the system. Similarly, the participant again

completed the generalized self-efficacy questionnaire so that we could measure whether self-efficacy increased after using the SIDNIE system.

The participant then completed a series of questionnaires on various characteristics of the system, including a System Usability Scale (Appendix I), a modified co-presence questionnaire, and a modified presence questionnaire (Appendix N). Finally, the participant completed a debriefing questionnaire (Appendix O) that provided qualitative feedback on the system, its usability, and perception of its effectiveness. The entire session lasted no more than one hour.

Two weeks later, the participants were e-mailed the same post-training questionnaire again so that we could measure learning retention.

8.2 Results

We recruited 62 first semester freshman nursing students enrolled at Clemson University School of Nursing. Inclusion criteria for participation in the research study included (1) English as first language, (2) self-reported 20/20 vision or 20/20 corrected vision, (3) hearing or corrected hearing within normal limits, and (4) 18 years of age or older.

Twelve participants were excluded due to technical difficulties or the inability to complete the experiment in the allotted time. Because there were repeated measures in this experiment (the learning outcomes and self-efficacy questionnaires), we excluded six additional participants because there was a missing questionnaire for their ID number, likely due to participant error in typing in their ID or a misunderstanding of what their ID number was. This left 44 participants' data for analysis.

8.2.1 Demographics

23 participants were in the non-scaffolded condition (Condition NS) while 21 participants were in the scaffolded condition (Condition S). There was no significant difference in the distribution of demographic qualities between the two conditions. The mean age of participants was 18.15, with a standard deviation of 0.37. There were three male participants and 41 female participants. 40 participants identified themselves as white, three participants identified as African or African American, and one participant was Hispanic/Latino.

We surveyed the participants about their experience levels with healthcare, children, com-

puters, and virtual reality. There were no significant differences in any of these experience levels between the conditions. See Figure 8.1 for a table of the reported experience levels.

Table 8.1: This table shows the average amount of experience participants reported in several categories, where 1 was labeled as "none" and 4 was labeled as "a lot".

	Mean	SD
Healthcare	1.90	0.77
Children	3.27	0.82
Children in healthcare	1.36	0.69
Computers	3.84	0.37
Virtual Humans	1.80	0.82
Virtual Reality	1.50	0.66

8.2.2 Learning Outcomes

Since the learning outcomes questionnaires were a repeated measure, we used a repeated measures MANOVA analysis. To measure effect size, we used $\eta^2_{partial}$ as well as Cohen's U^3 and Cohen's d. $\eta^2_{partial}$ were calculated using the method described in [Lakens, 2013], and Cohen's U^3 and Cohen's d were calculated and interpreted using the method described in [Magnusson, 2014].

8.2.2.1 Scoring Questions

In the learning outcomes questionnaire, students were asked to score ten questions on their sensitivity and language. We counted the number of questions each participant scored correctly for each attribute and analyzed the data to see if there was a significant difference between conditions or between the pre-test and post-test. See Figures 8.2.2.1 and 8.2.2.1 for boxplots of the overall scores as well as the scores separated by condition.

One participant's data was excluded from analysis due to being an outlier. For the sensitivity criteria, there was no significant difference between conditions ($F(1,41) = 3.85, p = 0.06, \eta^2_{partial} = 0.08$), between the pre-and post-questionnaires (F(1,41)=2.74, p=0.11), or interaction effect between the time of the questionnaire and the condition ($F(1,41) = 0.03, p = 0.87, \eta^2_{partial} < 0.01$). For the language criteria, there was no significant difference between the conditions ($F(1,41) = 1.12, p = 0.29, \eta^2_{partial} = 0.03$) or interaction between the time and the condition ($F(1,41) = 1.39, p = 0.24, \eta^2_{partial} = 0.03$); however, there was a significant difference between the pretest scores and the post-test scores regardless of condition ($F(1,41) = 8.21, p¡0.01, \eta^2_{partial} = 0.17$). Pre-test scores

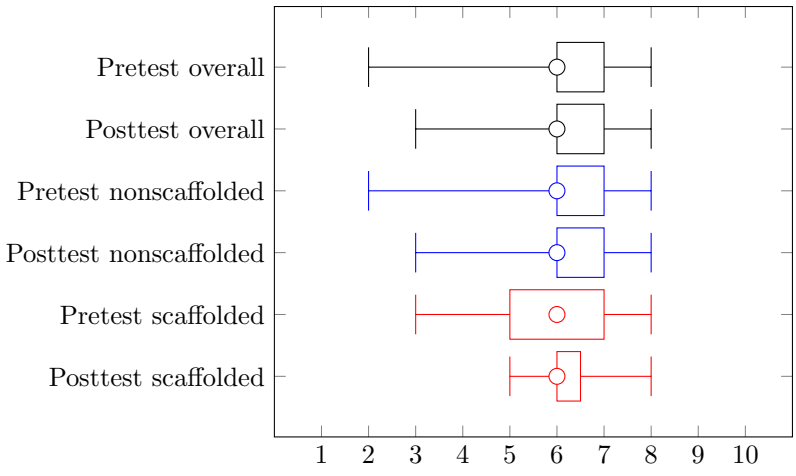


Figure 8.1: Boxplots for the pretest and posttest questions participants scored correctly for sensitivity. Whiskers show the minimum and maximum values. The median is indicated by a circle.

(mean=6.06, sd=1.07) were lower than post-test scores (mean=6.37, sd=0.95).

8.2.2.2 Writing Questions

The learning outcomes questionnaire also asked participants to write ten questions that they would ask the patient that were both sensitive and had appropriate language for the scenario. We submitted the resulting questions to Amazon Mechanical Turk and asked workers with a Master Qualification (above a certain acceptability threshold on previous tasks on Mechanical Turk) to score each question on both criteria. For this analysis, we excluded an additional three participants who were unable to complete the postquestionnaire due to time constraints, yielding 41 participants (21 in the nonscaffolded condition, and 20 in the scaffolded condition). Figures 8.2.2.3 and 8.2.2.3 show boxplots representing the distributions of participant scores according to their conditions.

The data for the pre-test and the post-test scores for language appropriateness met all assumptions for the MANOVA tests except that each data set was right-skewed from the normal distribution. However, the F-test is robust to departures from normality due to skewness. We ran a repeated measures MANOVA to determine whether there was any effect of time or condition. The analysis showed no significant difference between the two conditions ($F(1, 39) = 0.06, p = 0.81, \eta^2_{partial} < 0.01$). Although there was not a statistically significant difference, Cohen's d was d=0.63, indicating what is traditionally considered a "medium" effect size, and Cohen's U^3 was 74%, which indicates that 74% of the scaffolded group had a higher mean increase in questionnaire score

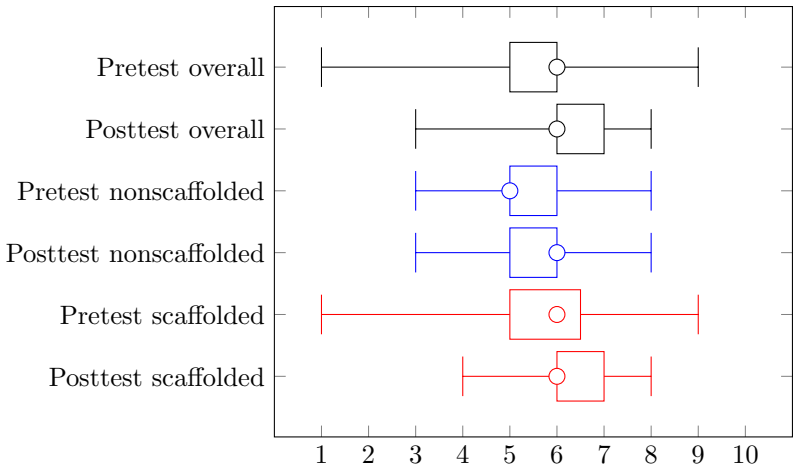
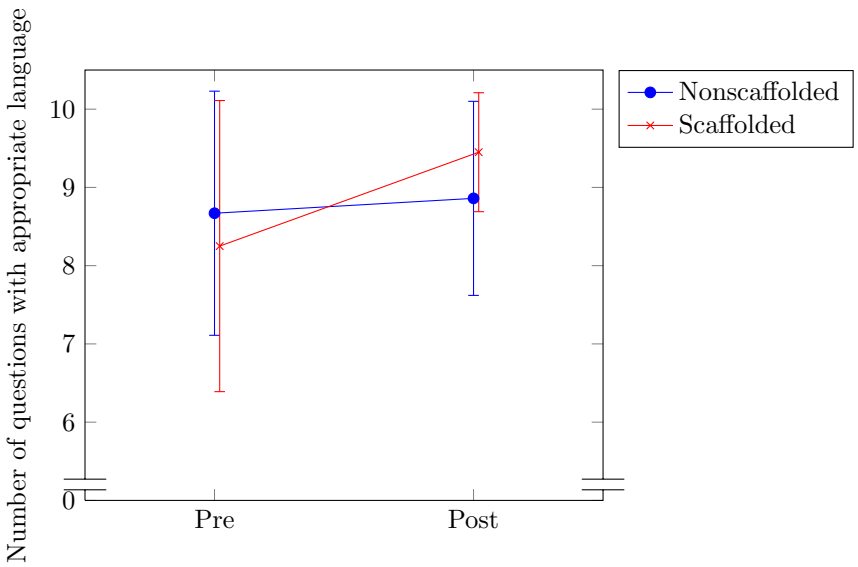


Figure 8.2: Boxplots for the pretest and posttest questions participants scored correctly for appropriate language. Whiskers show the minimum and maximum values. The median is indicated by a circle.

than the nonscaffolded group. In contrast, the $\eta^2_{partial}$ value was small, showing that an insignifcant amount of the variance between participants was due to their assigned condition.

However, there was a significant effect of time ($F(1, 39) = 7.54, p < 0.01, \eta^2_{partial} = 0.16$), where participants scored significantly higher in the postquestionnaire (mean=9.14, sd=1.70) than in the prequestionnaire (mean=8.46, sd=1.06). A closer look at the analysis shows that there was also a near-significant interaction effect between time and condition ($F(1, 39) = 3.97, p = 0.053, \eta^2_{partial} = 0.09$), where participants in the scaffolded condition tended to score lower on the prequestionnaire and higher on the postquestionnaire than those in the non-scaffolded condition, and the $\eta^2_{partial}$ value of 0.09 indicates that 9% of the variance was due to this interaction effect. Please see Figure 8.2.2.2 for a graph of the results showing the interaction between the conditions as well as the means and standard deviations of each condition.

We ran a similar analysis for the scores for sensitivity. We excluded one participant due to their postquestionnaire score being an outlier, resulting in 21 participants in the non-scaffolded condition and 19 participants in the scaffolded condition. Again, the pre- and post-questionnaire data departed from normality while all other assumptions were satisfied. The analysis showed a non-significant difference between the two conditions ($F(1, 38) = 3.79, p = 0.059, \eta^2_{partial} = 0.09$), where participants in the scaffolded condition showed a greater score increase than those in the nonscaffolded condition. Although the difference was not quite significant, Cohen's d was d=0.70, indicating what ist traditionally considered a "medium" effect, and Cohen's U^3 was 76%, which



	Pretest		Posttest	
	Mean	SD	Mean	SD
Nonscaffolded	8.66	1.56	8.86	1.24
Scaffolded	8.25	1.86	9.45	0.76

Figure 8.3: Questions that participants wrote in their prequestionnaire and postquestionnaire, scored as to whether they showed appropriate language. Error bars show one standard deviation.

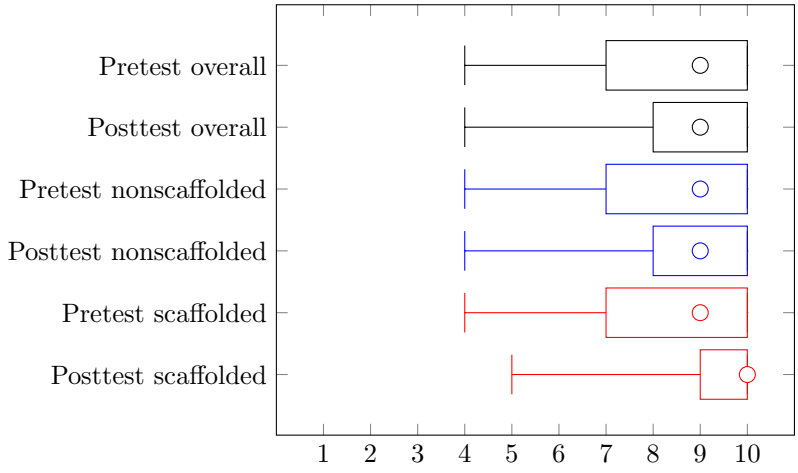
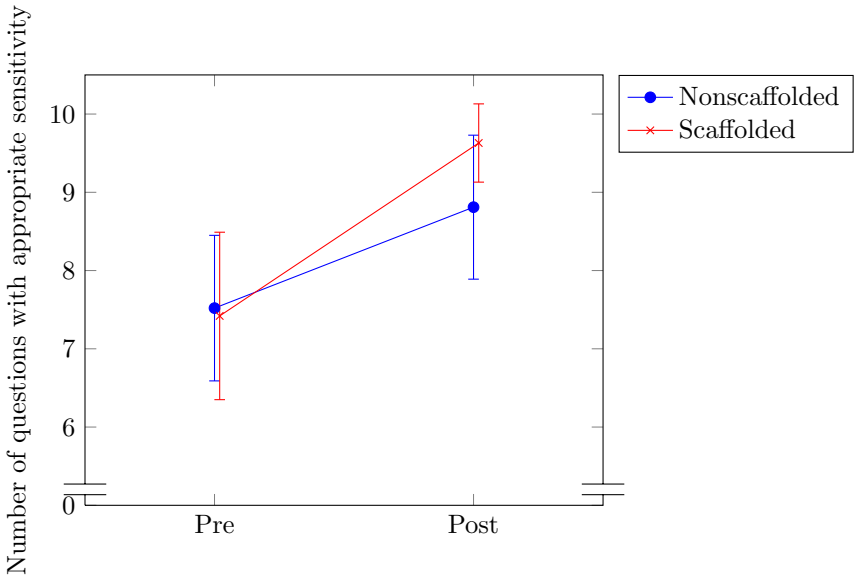


Figure 8.4: Boxplots for the pretest and posttest questions participants wrote, scored for appropriate language. Whiskers show the minimum and maximum values. The median is indicated by a circle.



	Pretest		Posttest	
	Mean	SD	Mean	SD
Nonscaffolded	7.53	0.93	8.81	0.92
Scaffolded	7.42	1.07	9.63	0.50

Figure 8.5: Questions that participants wrote in their prequestionnaire and postquestionnaire, scored as to whether they showed sensitivity. Error bars show one standard deviation.

indicates that 76% of the scaffolded group had a higher mean increase in questionnaire score than the nonscaffolded group. Additionally, the $\eta^2_{partial}$ value of 0.09 indicates that 9% of the variance was due to the participant's condition.

There was also a significant difference between the overall pre-test and post-test scores ($F(1, 38) = 68.70, p < 0.01, \eta^2_{partial} < 0.01$). Finally, there was a significant interaction between the time and the condition ($F(1, 38) = 4.81, p = 0.03, \eta^2_{partial} = 0.11$), where participants in the scaffolded condition again scored lower on the pre-test and higher on the post-test than those in the nonscaffolded condition, with this interaction accounting for 11% of the variance between participants. Please see Figure 8.2.2.2 for a graph of the results showing the interaction between the conditions as well as the means and standard deviations of each condition.

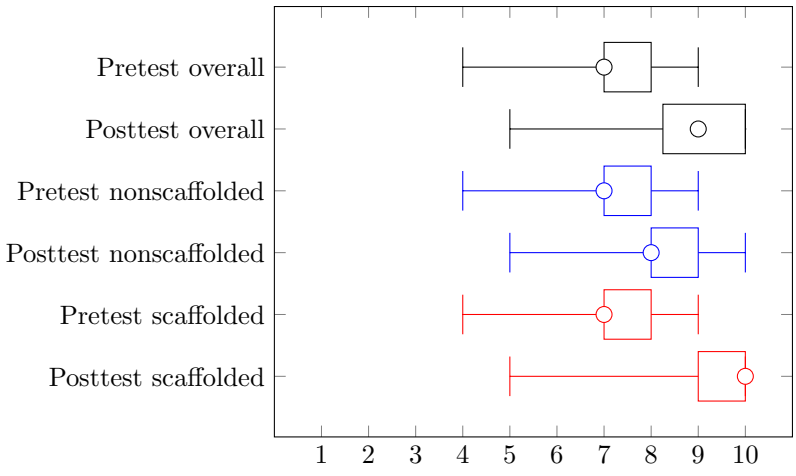


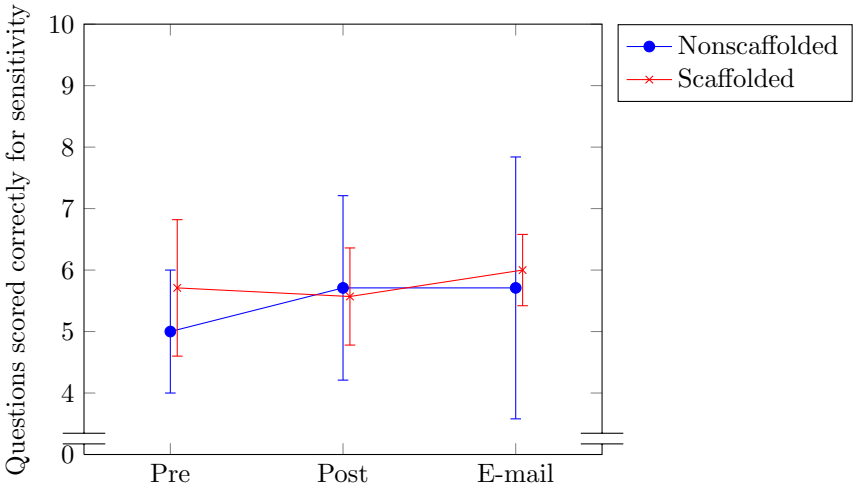
Figure 8.6: Boxplots for the pretest and posttest questions participants wrote, scored for sensitivity. Whiskers show the minimum and maximum values. The median is indicated by a circle.

8.2.2.3 Learning Retention

Two weeks after participation, each participant was e-mailed a link for a final questionnaire–one more repeated measure of the learning questionnaire. Only 22 participants chose to complete the final questionnaire, and eight of those participants were excluded due to a missing pre- or post-questionnaire or incorrect ID number, leaving 14 participants' data, with seven participants in each condition. We used a repeated measures MANOVA for this analysis paired with $\eta^2_{partial}$ to measure effect size.

Scoring Questions For the scored questions in the sensitivity criteria, there was no significant effect of condition ($F(1,12) = 0.0047), p = 0.95, \eta^2_{partial} < 0.01$) or time ($F(2,11) = 0.44, p = 0.65, \eta^2_{partial} = 0.07$), or interaction between condition and time ($F(2,11) = 0.389, p = 0.69, \eta^2_{partial} = 0.06$). See Figure 8.2.2.3 for a graph and chart of the mean values by condition and testing time for sensitivity. For the criteria of language, there were also no significant effects of condition ($F(1,12) = 0.21, p = 0.65, \eta^2_{partial} = 0.02$), time ($F(2,11) = 1.14, p = 0.35, \eta^2_{partial} = 0.17$), or interactions between condition and time ($F(1,12) = 1.74, p = 0.22, \eta^2_{partial} = 0.13$). See Figure 8.2.2.3 for a graph and chart of the mean values by condition and testing time for language. Also see Figures 8.2.2.3 and 8.2.2.3 for boxplots representing the distribution of responses.

Writing Questions For the written questions, three additional participants were excluded due to a lack of a complete questionnaire, yielding six participants in each condition. The questions



	Pretest		Posttest		E-mail	
	Mean	SD	Mean	SD	Mean	SD
Nonscaffolded	5.00	1.00	5.71	1.50	5.71	2.13
Scaffolded	5.71	1.11	5.57	0.79	6.00	0.58

Figure 8.7: Questions that participants scored correctly for sensitivity. Error bars show one standard deviation.

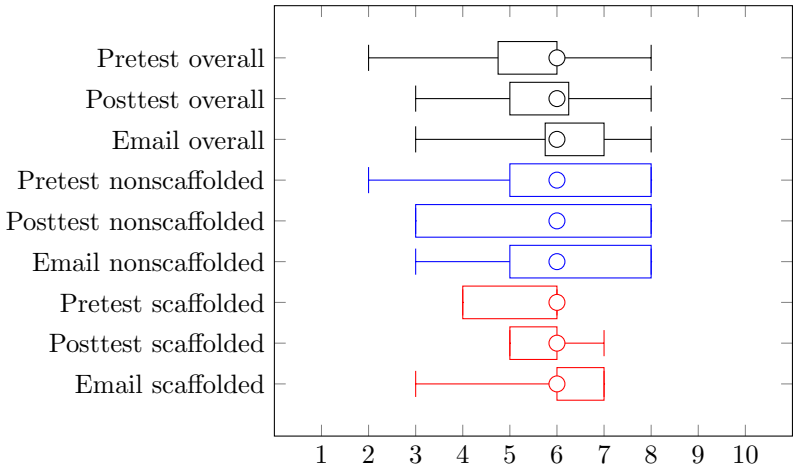
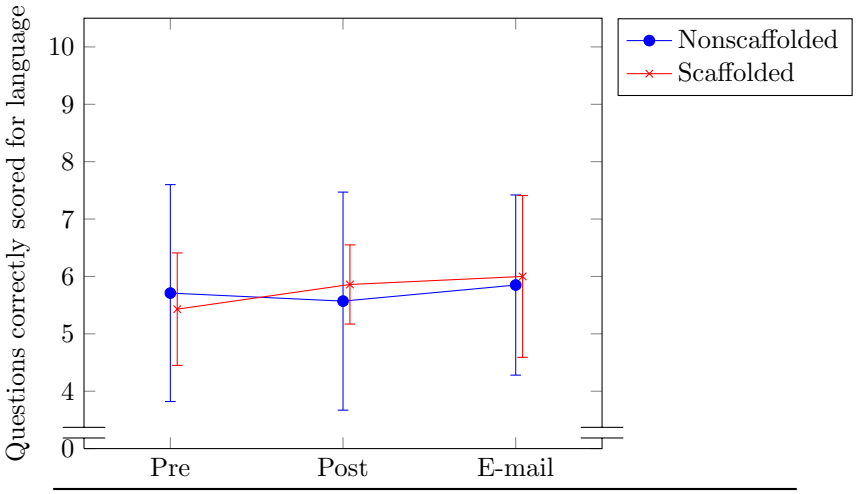


Figure 8.8: Boxplots for the pretest, posttest, and e-mail questionnaires where participants scored questions for appropriate sensitivity. Whiskers show the minimum and maximum values. The median is indicated by a circle.



	Pretest		Posttest		E-mail	
	Mean	SD	Mean	SD	Mean	SD
Nonscaffolded	5.71	1.89	5.57	1.90	5.86	1.57
Scaffolded	5.42	0.98	5.85	0.69	6.00	1.41

Figure 8.9: Questions that participants scored correctly for language. Error bars show one standard deviation.

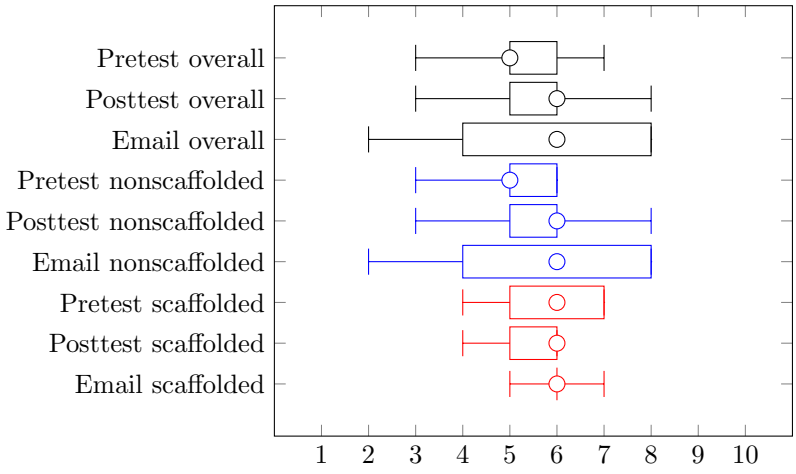
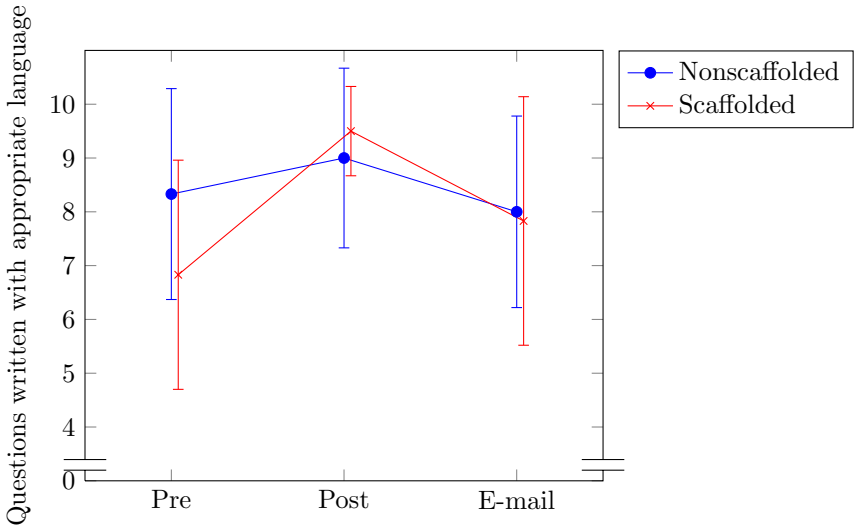


Figure 8.10: Boxplots for the pretest, posttest, and e-mail questionnaires where participants scored questions for appropriate language. Whiskers show the minimum and maximum values. The median is indicated by a circle.



	Pretest		Posttest		E-mail	
	Mean	SD	Mean	SD	Mean	SD
Nonscaffolded	8.33	1.96	9.00	1.67	8.00	1.78
Scaffolded	6.83	2.13	9.50	0.83	7.83	2.31

Figure 8.11: Questions that participants wrote that had appropriate language. Error bars show one standard deviation.

generated by those twelve participants were scored using Amazon Mechanical Turk for their language and sensitivity. For language, there was no significant difference between the two conditions ($F(1,10) = 0.17, p = 0.69, \eta^2_{partial} = 0.02$) or interaction between time and condition ($F(2,9) = 2.19, p = 0.16, \eta^2_{partial} = 0.18$); however, there was a significant effect of time ($F(2,9) = 8.58, p < 0.01, \eta^2_{partial} = 0.65$), with the $\eta^2_{partial}$ statistic suggesting that 65% of the variance in scores was due to the factor of time. After running pairwise t-tests between the three questionnaires' average scores, we found that the post-questionnaire score was significantly higher than both the prequestionnaire score ($t(11) = 3.15, p > t = 0.0045$) and the questionnaire given via e-mail two weeks later ($t(11) = 3.54, p > t = 0.0023$), while there was no significant difference between the pre-test and the questionnaire given over e-mail two weeks later ($t(11) = 0.67, p > |t| = 0.51$).

Similarly, for sensitivity, there was no significant difference between the two conditions ($F(1,10) = 2.42, p = 0.15, \eta^2_{partial} = 0.19$), and no interaction effect between time and condition ($F(2,9) = 2.070, p = 0.12, \eta^2_{partial} = 0.31$), but a significant effect of time ($F(2,9) = 39.58, p < 0.01, \eta^2_{partial} = 0.90$), with , $\eta^2_{partial}$ indicating that the time differences could account for 90% of the differences in variance. The pairwise t-tests show that there were significant differences in all three

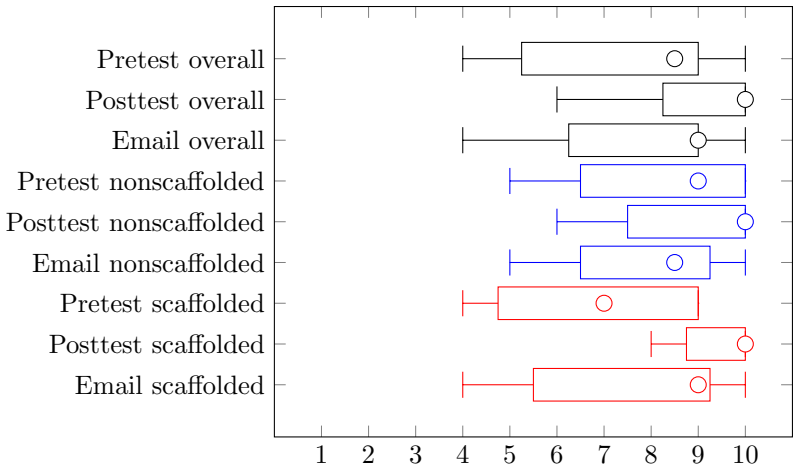


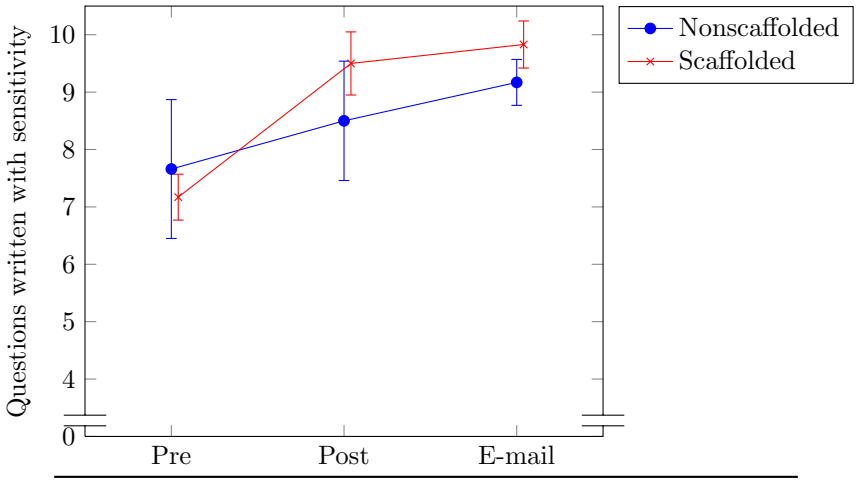
Figure 8.12: Boxplots for the pretest, posttest, and e-mail questionnaire scores where participants wrote questions with appropriate language. Whiskers show the minimum and maximum values. The median is indicated by a circle.

pairings, where the postquestionnaire scores were higher than the prequestionnaire scores ($t(11) = 3.64, p > t = 0.0019$), and the e-mail questionnaire scores were higher than both the prequestionnaire scores ($t(11) = 7.24, p > t < 0.001$) and the postquestionnaire scores ($t(11) = 1.91, p > t = 0.04$).

8.2.3 Self-Efficacy

The self-efficacy score was calculated by summing together all the scores for the self-efficacy questionnaire, yielding scores ranging from 8 (lowest self-efficacy) to 40 (highest self-efficacy). One participant was excluded from analysis due to his or her prequestionnaire self-efficacy score being an outlier. Using a repeated measures MANOVA, we found no significant difference in self-efficacy scores overall between conditions ($F(1, 41) = 0.404, p = 0.84, \eta^2_{partial} < 0.01$). There was also no significant difference in the pre-test self efficacy scores and the post-test self efficacy scores ($F(1, 41) = 1.14, p = 0.29, \eta^2_{partial} = 0.02$), or interaction effects between the condition and the time ($F(1, 42) = 0.11, p = 0.74, \eta^2_{partial} = 0.02$). Overall the self-efficacy score mean for the pretest was 33.79 ($sd = 2.45$), and the mean for the post-test was 34 ($sd = 2.79$).

We also analyzed each question of the self-efficacy questionnaire to see if there were any significant differences between the conditions, or between the pre- and post-test questionnaires. For the statement "When facing difficult tasks, I am certain that I will accomplish them," while there was no effect of condition or interaction between condition and time, there was a significant difference



	Pretest		Posttest		E-mail	
	Mean	SD	Mean	SD	Mean	SD
Nonscaffolded	7.66	1.21	8.50	1.04	9.17	0.41
Scaffolded	7.17	0.41	9.50	0.55	9.83	0.41

Figure 8.13: Questions that participants wrote that had appropriate sensitivity. Error bars show one standard deviation.

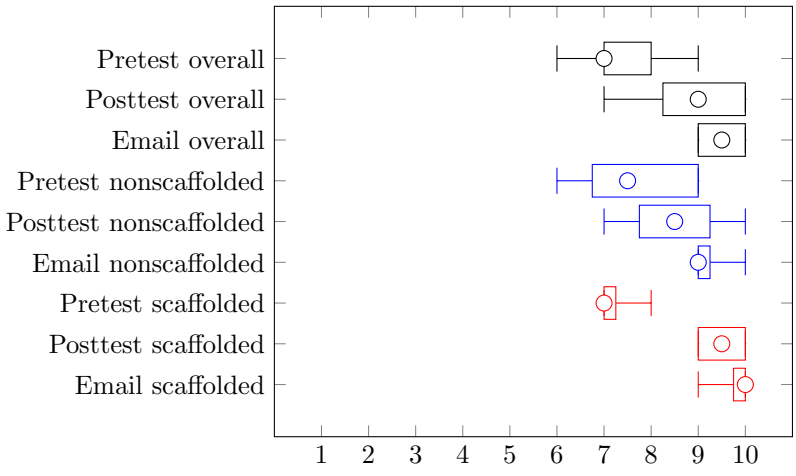


Figure 8.14: Boxplots for the pretest, posttest, and e-mail questionnaire scores where participants wrote questions with appropriate sensitivity. Whiskers show the minimum and maximum values. The median is indicated by a circle.

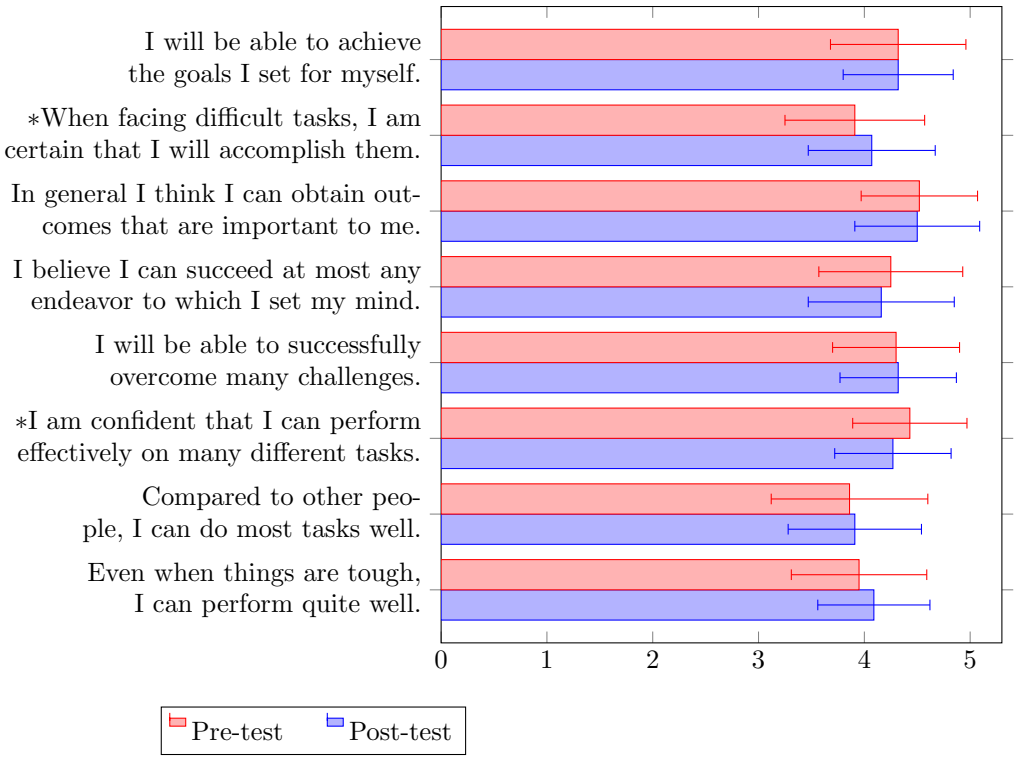


Figure 8.15: Results of the self-efficacy questionnaire comparing pre-test and post-test scores. Error bars show one standard deviation in each direction. Questions labeled with a $*$ showed a significant change between pretest and posttest scores.

between pretest scores and post-test scores ($F(1, 42) = 2.23, p = 0.04, \eta^2_{partial} = 0.05$), with pretest scores ($mean = 3.91, sd = 0.60$) being lower than post-test scores($mean = 4.07, sd = 0.66$). For the question "I am confident that I can perform effectively on many different tasks," there was a significant difference between pre-test and post-test scores ($F(1, 42) = 4.68, p = 0.04, \eta^2_{partial} = 0.10$), although in this case scores in the pre-test ($mean = 4.43, sd = 0.55$) were higher than those in the post-test ($mean = 4.27, sd = 0.54$). Please see Figure 8.15 for a bar chart of means and standard deviations for each question.

8.2.4 System Usability

The overall System Usability Score was calculated by reverse-scoring the reversed items in the questionnaire, summing together the scores for each of the 10 items in the measure, then dividing by 70 and multiplying by 100 for a score between 0 (lowest usability) and 100 (highest usability). Because this measure was not repeated, we used t-tests to analyze results, and η^2 to measure effect

size.

Between conditions, the SUS score variances were not significantly different, so using a t-test assuming equal variances, we found no significant differences between the overall usability scores ($t(42) = 1.25, p = 0.2175, \eta^2 = 0.04$). The mean usability score was 85%, with a standard deviation of 7.97. This shows that overall participants found the SIDNIE system usable.

We also gave the participants the option to make any comments about the system's usability in a text box. Six participants had no comments. 26 participants out of the 44 specifically mentioned that they thought the system was easy to use. Six participants commented that they thought SIDNIE would be a helpful tool for themselves or other medical professionals to begin developing their communication skills. In the non-scaffolded condition, one participant requested a better tutorial before starting the system, and another participant expressed confusion about whether to select multiple questions and how that would affect his or her summative scoring in the end. In the scaffolded condition, one participant expressed support for the scaffolded model: "SIDNIE is easy to use as it has detailed instructions that move at the pace of the person using the program. It allows the user time to process what they are doing and learn as they go. The set-up of having the user listen to the simulation the first time, have a guided interview the second time, and then finish with an interview without guidance is perfect for those who are learning. They get to build up the skills at a good pace through this application."

Participants expressed some usability concerns. Due to a technical error, occasionally in the observed interview, the patients responded to the question before SIDNIE asked it. A couple participants mentioned this shortcoming. The other salient usability problem had to do with the voices used for text-to-speech. Four students mentioned difficulty differentiating between the voices and which character they belonged to. We only had a limited number of text to speech voices on the system, several of which sounded similar.

We analyzed each question in the SUS questionnaire to see if there were significant differences between conditions. Only one question showed a significant difference between conditions: "I needed to learn a lot of material before I could get started with SIDNIE." The response did not come from a normal distribution and the variances between the conditions were not equal, so instead of a t-test we used a Kruskal-Wallis test to detect the difference ($\chi^2 = 3.89, p = 0.049$). Participants in the non-scaffolded condition ($mean = 3.00, sd = 1.78$) agreed more highly with the question than those in the scaffolded condition ($mean = 1.90, sd = 0.99$).

8.2.5 Presence and Copresence

First, we calculated an overall presence and copresence score, where we counted each of the sixteen questions that a participant scored with 4 or higher out of 7. A t-test showed no significant differences in those scores between conditions ($t(42) = -0.374, p > |t| = 0.7103, \eta^2 < 0.01$). The mean count score across all conditions was 7.04 ($sd = 3.98$) out of 16 questions, showing that participants did at least have some sense of "being there". Scores also ranged from 0 to 14, showing that even though there were not differences between conditions, there were differences in how individuals perceived the experience.

We analyzed each question separately to determine whether any individual question had significant differences between conditions. Only one question showed a significant difference. For the question "To what extent did the virtual patients help you in carrying out the task?", there was a significant difference between conditions ($t(42) = -1.79, p < t = 0.041, \eta^2 = 0.07$), with participants in the non-scaffolded condition ($mean = 5.13, sd = 1.21$) finding the patients more helpful than those in the scaffolded condition ($mean = 4.29, sd = 1.87$). To see the means and standard deviations for each question in the presence and copresence questionnaire, please refer to Figure 8.16.

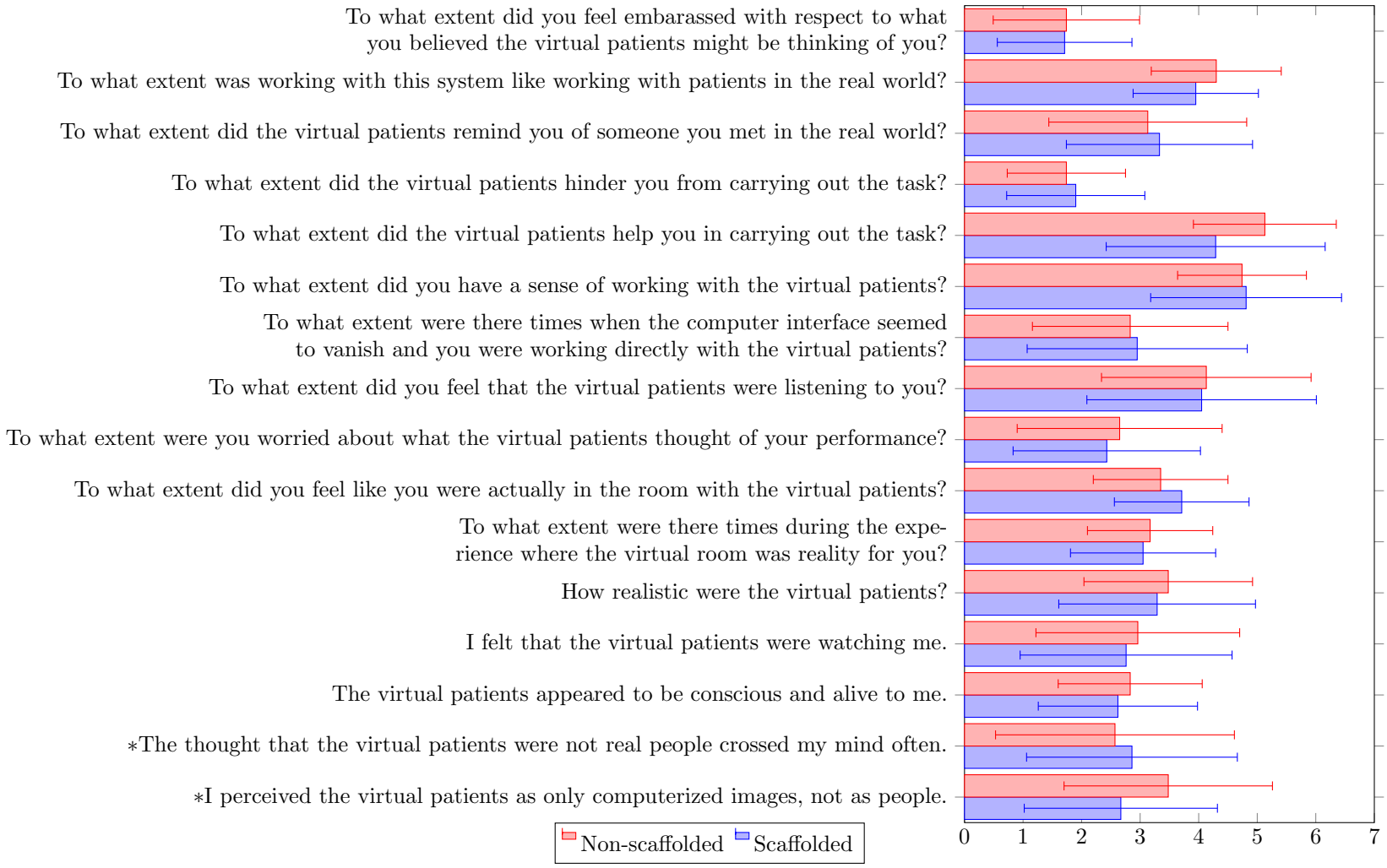


Figure 8.16: This chart shows the means and standard deviations for the presence and copresence scores divided by condition. Questions with scoring reversed are indicated with a *. The error bars represent one standard deviation in each direction.

8.2.6 Subjective Feedback

At the end of the post-questionnaire, participants had the option to give free-text feedback for nine questions asking their general impressions about the system. The first question sought to gather general feedback about the system by asking, "What did you think about using this system?" 41 of the 44 participants noted positive aspects of the simulation. Two participants made neutral statements ("it was decent","it was okay"). One participant said, "I personally do not like virtual patients or any sort of technological interface. I prefer face to face interaction." Seven participants said they thought the patients were not realistic, while three participants said they thought the patients were realistic. Twenty participants said that they thought it would be useful for practice or learning how to communicate with actual patients. One participant said, "It allows a student to first begin learning about patient interaction in a no stress environment." Twelve participants expressed that they enjoyed using the system, that it was "cool," "innovative," "eye-opening" or "unique." Eight participants said that the system was easy to use.

When asked what they liked the best about interacting with SIDNIE, responses varied greatly. Six participants listed some aspect of the characters' realism (that they were breathing, that they had human faces, etc.). Thirteen participants liked that they were given multiple questions to select from to ask the patient. Four participants liked the interaction without consequences, saying that "I was not worried about doing something incorrectly," "the people do not respond with anger," and "I didn't feel like the patients were judging me." Four participants said their favorite aspect was interacting with a variety of virtual patients in the simulation. Among the participants in the scaffolded condition, six participants mentioned some aspect of the scaffolding as what they liked the best–SIDNIE's pacing, the ability to receive immediate feedback on their question choices, or the modeling of the interview process by SIDNIE.

When asked what they disliked the most about interacting with SIDNIE, three participants found the question choices difficult. One participant was uncomfortable during the simulation, saying that "staring at a huge screen so up close makes my vision fuzzy with contacts." We also directly asked whether the participant thought the patients were realistic. 32 participants cited some aspect of SIDNIE's lack of realism in terms of either how the characters looked, move, or spoke. Participants found the appearance and the text to speech voices "creepy" and "disconcerting". Nine participants agreed that they were realistic while the remaining 36 participants said at least one

aspect of the patient was unrealistic, primarily their appearance or the sound of their voices.

We also asked two questions about learning outcomes. First, we asked, "How well do you think you learned from this system?" Only one participant expressed a lack of learning, saying "I do not think I learned well at all." The remaining 43 participants expressed at least some level of perceived learning. Seven participants said they thought they had learned "what types of questions to ask a patient". Seven participants mentioned learning some aspect of correct "word choice". One participant said, "I think it was better than simply reading how to interact with patients in a text book." Another participant said, "It is a good opportunity to practice social interaction with minimal consequences." We then asked if there was anything that could be changed about SIDNIE to better support learning. 22 participants did not give any suggestions. Four participants suggested better feedback about why a selection was correct or incorrect. Eight participants asked for "more question freedom" or the ability to write or speak their own questions. Four participants wanted a greater variety of scenarios or a greater variety of content within scenarios. Two participants wanted the patients to be more realistic. The remaining participants cited problems with understanding the directions or terminology.

When asked whether they thought the patient scenarios were realistic, all but one participant responded that they considered the scenario to be realistic. We also asked whether the participant would choose to use the system again for practice. All but four participants answered they would, while the remaining participants said they "possibly" would. One participant commented that using the system would "be great practice for myself and building up my confidence." Finally, we asked whether the participant thought using SIDNIE would affect their clinical experiences. All but one participant said they thought it would. Participants were aware of the system's limitations, saying that "the patients are not real...nurses will often interact with people who...have emotions" and "I think I would learn better by following a real nurse...this could be a great supplement to clinical experiences." However, they felt that they could gain many skills from the simulation, saying that it would "make me more confident," "help me better connect with the patients," "be more prepared for when you actually go into clinicals," and "take away some of the nervousness of working with patients."

8.3 Discussion

When interpreting the results of this experiment, it is important to remember that the scenarios that the participants experienced using SIDNIE took little effort or medical expertise to create using the scenario builder tool, and that the patient's nonverbal behaviors were intentionally as simple as possible. Each scenario was created in less than 30 minutes and relied heavily on readily-found case studies online, and automatically generated questions and answers. Given the opportunity, a nursing instructor could make scenarios that much more closely resemble actual clinical interaction, and could tailor the content better to a specific case study and desired learning goal.

There were no nonverbal cues except for breathing and lipsyncing animations, and all patient responses were text-to-speech. This may account for the vast majority of the complaints about usability, lack of presence, and lack of realism. Allowing nurse educators to record speech instead of using text-to-speech would be a significant technical challenge (as well as a time-consuming task for the nurse educator). However, text-to-speech voices are constantly improving, and using higher quality voices is possible. Additionally, adding symptom-linked animations is a possible avenue for future work on this project. Participants certainly would prefer better animations and better voice quality, and so it is worthwhile future work to pursue in the interest of user satisfaction. However, as the results of this experiment confirm (as well as our previous fidelity experiments originally showed in [Pence et al., 2013]), these higher-fidelity attributes are unnecessary for learning verbal skills, and any future development should balance the cost in terms of development time and nurse educator time against the expected learning and usability gains for each feature.

Only one scaffolding configuration was represented in this experiment, with participants moving sequentially and immediately between observing, participating in a guided interview with immediate feedback, then participating in an interview with little guidance and only summative feedback. There may exist a more ideal scaffolding structure with more levels or different feedback structures, although the analysis suggests that this scaffolding sequence is at least somewhat functional. Additionally, if students were required to practice each level until achieving competence before moving on to the next level, or to move through levels over multiple sessions, learning outcomes may improve. This experiment was designed to test the feasibility of such a system but in the classroom would likely be best used differently.

8.3.1 Learning Outcomes

We measured learning outcomes through a repeated measures questionnaire in two ways: (1) having participants score ten questions according to the criteria, and (2) having participants write ten questions that met the criteria. These tasks were selected because they are representative tasks of the "evaluating" and "creating" levels of Bloom's taxonomy, respectively. Each participant was invited to complete the questionnaire three times: immediately before using the SIDNIE system, immediately after using the SIDNIE system, and then an e-mail questionnaire two weeks after their participation in the experiment.

Our first hypothesis was that *participants who were in the scaffolded condition would show better learning outcomes than those in the non-scaffolded condition immediately after participation.* In the first measure, scoring questions, the hypothesis was not confirmed. There was no difference between the pre-test and post-tests scores due to the participant's condition. For the language criteria, however, there was a significant difference in pre-test and post-test scores regardless of condition, showing that with or without scaffolding, exposure to the SIDNIE system did help participants learn about appropriate language.

For the second learning measure, the participant-written questions, the data is much more promising. Regardless of their condition, participants learned, as there was a significant difference between pre-test and post-test scores for both criteria. When comparing results between conditions, although the F-tests fell short of significance, the Cohen's d calculations suggest that the condition had a medium effect size for both sensitivity and language, with participants in the scaffolded condition outperforming participants in the nonscaffolded condition in each case. Additionally, there were interaction effects in each criteria that are consistent with the scaffolded condition being superior, since for each criteria, participants in the scaffolded condition scored lower on average in their pre-test and higher on average in their post-test than those in the nonscaffolded condition.

Several of the F-tests for the participant-written questions were very close to significant values, so it is possible that with more participant data a statistically significant effect could be uncovered. Nevertheless, statistical significance and practical significance are not always equivalent. From a practical standpoint, the data shows that on average, the scaffolded participants either outperformed or did as well as their nonscaffolded counterparts in all measures in this experiment. Additionally, from a qualitative and usability perspective, at least six participants specifically named

something about SIDNIE's scaffolding as their favorite part of the simulation.

My second hypothesis was that *participants who were in the scaffolded condition would show better learning retention two weeks after participation than those in the non-scaffolded condition.* The data collected in this experiment does not support this hypothesis. In both the scored questions and the written questions, there was no significant difference due to condition. There was a very low response rate for the learning retention measure, with only 23 participants responding, and only 13 of those participants yielding a complete pre-test, post-test, and e-mail test for analysis. It is possible that new results would emerge with the collection of more data. It is also possible that the scaffolding structure or the time between scaffolding sessions should be adjusted to maximize learning retention.

8.3.2 Self Efficacy

My final hypothesis was that *participants who were in the scaffolded condition would show a greater increase in self efficacy than those in the non-scaffolded condition.* This hypothesis was not supported by the data we collected, since there was no significant difference in self-efficacy scores between conditions or overall between the pre-test and the post-test, regardless of condition. Two questions did show a difference between post-test and pre-test regardless of condition. Participants ranked "When facing difficult tasks, I am certain that I will accomplish them" lower on the pre-test than the post-test, showing a possible increase in that aspect of self-efficacy due to their interaction with the SIDNIE system. Although there is no clear explanation for the increase, it could be that participants were somewhat lacking in confidence during the pre-test due to being in an unfamiliar environment and being uncertain about the task to come, while after interaction with SIDNIE they were less anxious.

However, participants rated a second question, "I am confident that I can perform effectively on many different tasks", lower in the post-test than in the pre-test, showing a decrease in that aspect of self-efficacy due to interaction with the SIDNIE system. Again, there are no clear explanations for this change in sentiment. It is possible that students felt less confident in their abilities after seeing their scores for each scenario they interacted with.

Although there was not clear statistical evidence for SIDNIE's effectiveness in increasing self-efficacy, several user comments suggest that it could be useful for increasing confidence. At least seven participant comments mentioned in some form that they thought that SIDNIE gave them the

practice they needed to build confidence and practice interactions without any pressure from actual patients, and 43 out of the 44 participants reported that they felt like they had learned something during their interaction with SIDNIE, and that it would positively affect their clinical experiences.

Overall, self-efficacy scores were high, with the pre-test average being 33.79 out of 40. The self-efficacy scale we chose was not specific to any context, so it measured a general sense of self efficacy. In future experiments, it may be more informative to use a self-efficacy scale that has been rephrased for a nursing context (for example, question such as "I am confident that I can perform effectively on many different tasks during my clinical experiences") to better measure situational confidence. Freshmen nursing students are likely not as confident in their nursing abilities as they are in their ability to succeed in life in general, so there may be more room for change between pre-test and post-test in that context.

8.4 Comparing Results with Other Studies

Because the studies presented in Chapter 5 used some of the same subjective measures, we can also informally consider whether a scenario created using the traditional methods yielded different subjective results than a scenario created using the scenario builder tool. The comparison must be considered cautiously, since the participants were from different audiences and the tasks were somewhat different; however, comparing the scores may give insight into differences in the way participants perceived the scenarios, since interaction with the SIDNIE system itself changed only minimally (as described in Section 5.3).

8.4.1 Presence

The study comparing various fidelity levels in Section 5.2 used many of the same presence and copresence questionnaires as this study. We ran two-sided t-tests to determine whether there were any significant differences in presence scores between the two studies, considering the study as the only factor. Considering the presence score overall (the count of items with a score greater than 4), after removing two outliers (leaving 20 participants' data in the fidelity experiment, and 43 participants' data in the second experiment), there was a significant difference in the presence scores ($t(61) = 1.90, p = 0.03, \eta^2 = 0.06$), with the presence scores in this scaffolding evaluation (mean=4.35, sd=2.89) higher than the presence scores in the fidelity study (mean=3.00, sd=1.91).

Considering the results question by question, the differences found to be insignificant were equally interesting as those found to be significant. For the question "To what extent, if at all, were you worried about what the virtual patients thought of your performance?" after removing an outlier and using a t-test assuming unequal variances, there was a significant difference ($t(62) = 4.33, p < 0.01, \eta^2 = 0.15$), with participants in the scaffolded study (mean=2.46, sd=1.59) scoring the question higher than those in the fidelity study (mean=1.28, sd=0.56). This difference could be due to the differences in task. In the fidelity study, the participants interacted with one scaffolding level, while in the scaffolding study, the participants interacted with three levels, receiving more feedback. For the similar question "To what extent did you feel embarrassed with respect to what you believed the virtual patients might be thinking about you?", the statistic should be interpreted with some caution since the response distribution departed from normality, but a two-sided t-test assuming unequal variances showed that there was no significant difference between the two experiments, after excluding one outlier ($t(62) = 1.38, p = 0.17, \eta^2 = 0.02$).

There were no significant differences for the two questions concerning the "relatability" of the characters ("To what extent did the virtual patients remind you of someone you met in the real world?" and "To what extent was working with this system like working with patients in the real world?"), suggesting that the characters created automatically through the scenario builder tool were equally "familiar" as the characters purchased for the first SIDNIE iteration. Similarly, the questions about whether the virtual characters helped or hindered the participant, and whether they had a sense of working with the virtual patient, showed no significant differences, as did the questions about whether the character appeared conscious and alive, or seemed to be listening or watching the participant.

However, when it came to the realism of the virtual patient, there was a significant difference ($t(63) = -1.88, p = 0.03, \eta^2 = 0.05$), with participants in the scaffolding study (mean=3.36, sd=1.54) rating the realism lower than those in the fidelity study (mean=4.09, sd=1.09). Since this experiment, the visual fidelity of the characters has been improved (as described in Section 7.1.1–despite the ordering in this document, the study in Chapter 7 was performed after this study); however, this result may indicate that there is still work to be done in improving realism. In contrast, the questions about whether the virtual characters were perceived as real people or as computerized images showed no significant differences between experiments.

For the question "To what extent were there times during the experience when the virtual

room was reality for you?", there was also a significant difference between the two experiments ($t(63) = 3.90, p < 0.01, \eta^2 = 0.19$). The scaffolding evaluation experiment scored higher (mean=3.11, sd=1.14) than the fidelity experiment (mean=1.95, sd=1.07). In contrast, there was no significant differences between experiments in the questions about feeling like you were actually in the virtual room, or whether the computer interface seemed to vanish.

8.4.2 System Usability Scale

The initial study on SIDNIE's learning performance in Section 5.1 also used the SUS scale to measure usability. In the first study, there were 15 participants' data, while in the second study there were 44 participants. Comparing the overall SUS results, there was a significant difference between the two usability scores ($t(57) = 1.69, p > t = 0.05, \eta^2 = 0.05$), with the mean usability score for the second experiment (mean=84.96%, sd=7.97) being higher than the usability score for the original experiment (mean=80.33%, sd=12.20).

There were four individual questions with significant differences. All questions showed unequal variances between the two experiments, so a t-test assuming unequal variances was used for analysis. After removing one outlier, the question "I thought the system easy to use" was scored higher in the current study (mean=5.20, sd=0.97) than in the first study (mean=4.13, sd=1.85, $t(56) = 2.15, p = 0.02, \eta^2 = 0.13$). This shows that either the changes in the scaffolding levels or in the scenario itself increased overall usability somewhat. Similarly, after removing two outliers, there was also a significant difference in scores for the question "I would imagine that most people would learn to use this system very quickly" ($t(55) = 2.09, p = 0.03, \eta^2 = 0.15$). Participants in this experiment agreed with the question more (mean=5.54, sd=0.59) than those in the first experiment (mean=4.66, sd=1.59). Following the same pattern, the question "I felt very confident using the system" showed higher scores ($t(56) = 2.11, p > t < 0.01, \eta^2 = 0.13$) in the current experiment (mean=4.93, sd=0.17) than in the original experiment (mean=3.93, sd=0.30).

The anomalous question was "I needed to learn a lot of things before I could get going with this system." When the scoring was reversed so that higher scores yield better usability, the first experiment (mean=5.60, sd=0.63) yielded better scores than the current experiment (mean=4.52, sd=1.54, $t(57) = -3.78, p < 0.01, \eta^2 = 0.11$). One possible explanation for the difference was the students' level. In the current experiment, students were second semester freshmen, while in the first experiment, the majority of participants were accelerated BSN students who already had obtained a

bachelor's degree in some other field. Older students might have brought more pre-existing knowledge to their experiences.

8.4.3 Discussion

Again, these comparative results should be interpreted with caution, due to differing audiences and differing tasks. However, the results do seem to indicate that in most aspects, using the scenarios and characters from the scenario builder tool in the current SIDNIE system is no less usable and causes no less of a sense of presence than the original SIDNIE system, and in some cases even increases usability and presence. Additionally, although no direct comparison is possible between the learning outcomes in the three studies, all three yielded positive learning outcomes, regardless of differences in scenario creation technique. Scenarios and characters created through the scenario builder interface are much less costly than creating scenarios and characters by hand (as in the first study). Since it seems that these less expensive techniques yield equally positive results, this suggests that the scenario builder tool is an appropriate and effective platform for future virtual patient development.

8.5 Recommendations for Other Applications

The vast majority of virtual reality training simulations in all domains implement pure experiential learning. The user is placed in a highly realistic virtual situation, then navigates through the scenario as best as he or she can. Often feedback is offered at only at the end of the scenario, or only implicitly by showing negative consequences for incorrect actions. Many of these simulations are effective in that they lead to learning gains.

However, in this study, the comparative evaluation shows that adding learning scaffolding that includes multiple levels (some of which give real-time, direct feedback), provides as good or better learning outcomes and usability results than a non-scaffolded presentation with feedback after interaction. Interaction interrupted by real-time feedback may not be realistic to the situation that the simulation represents. However, matching that aspect of realism seems unnecessary to foster learning. In fact, nearly every aspect of our scenario's realism was criticized by users (patients' appearance, voices, animations, questions and responses), yet users still showed positive learning outcomes. Creators of training scenarios in virtual reality should consider adding scaffolding for

learning within their simulations. Creators should also carefully consider and evaluate the level of realism that is necessary for learning to occur within their specific setting. It is possible that time spent increasing realism could better be spent increasing learning support to yield better educational outcomes.

Chapter 9

Conclusion and Future Directions

This work detailed the design and implementation of a scenario builder tool for a pediatric virtual patient system. Nurse educators were successful in using the scenario builder tool to generate scenarios for the SIDNIE system, and scenarios generated using the tool were shown to be effective for learning. The scenario builder tool reduced the workload for creating a scenario from approximately nine months to approximately a half-hour, with both the hand-encoded scenario and the scenario made with the tool showing positive results in terms of usability and learning outcomes. In summary, the contributions of this work to the field of computer science are:

- We contributed the SIDNIE (Scaffolded Interviews Developed by Nurses in Education) platform for virtual patient simulation. SIDNIE is novel in that it provides real-time scoring of student interaction and changeable learning criteria and learning scaffolding [Dukes et al., 2013]. User studies show that SIDNIE is usable and effective for learning in both low and high visual and interaction fidelity situations [Pence et al., 2013, Dukes et al., 2013].

- We contributed a scenario builder tool for SIDNIE that allows nursing educators to independently create their own virtual patient scenarios. Although the tool was designed specifically with and for nursing faculty, it is likely also be suitable for other novice users, and through its extensibility can be adapted to other conversational training tasks. Additionally, usability principles discovered through the participatory design and formal usability evaluation of the scenario builder tool may be applied to other situations where designers must make interfaces for novice users completing complex tasks.

- Through the development of the scenario builder tool, we contributed a pipeline to smoothly create, import, and animate new virtual characters in Unity3D at runtime, using the MakeHuman platform. This has the potential to simplify creation of many virtual character applications as well as to diversify the population groups represented in virtual human applications.
- We conducted a comparative evaluation of scaffolded and non-scaffolded presentations of virtual patient scenarios, finding that the scaffolded presentation of scenarios yielded learning outcomes that were as good as or better than those from a non-scaffolded presentation. This suggests that creators of future virtual environments for learning should consider adding in scaffolding for supporting learners as they complete complex tasks.
- We published several papers on the SIDNIE system, as listed in Appendix A.

9.1 Future Directions

The primary goal of this work was to create and evaluate a system that would allow nurse educators to extend the current SIDNIE system to represent scenarios of their own choosing. From this point there are several future research directions that could further this initial goal as well as answer new research questions about interaction with virtual environments and virtual characters.

9.1.1 Extensibility, Sharing, and Reuse

In the current implementation, the database for the scenario generation tool is locally stored on the computer where the application runs. This means that a nurse educator can build on his or her work over time; however, unless databases are copied between computers, other nurse educators cannot benefit from any outside contributed content. For future work, it would be possible to make this database available as a centralized online resource so that nurse educators could share their created assets, further increasing reuse and reducing workload.

However, with a centralized database comes many new challenges. With content being contributed from many different sources, to be effective, the content must be curated in many ways, including meaningful organization, searchability, metrics for quality, and the ability to find and remove duplicates. Additionally, a centralized database opens the door for collaboration, perhaps through an online platform, where multiple nurse educators can work to develop the same scenario,

or several similar scenarios can be merged to create a higher quality scenario. Scenarios and other assets could be shared and developed across institutions. Possibly, a standardized set of scenarios could be collaboratively developed for baccalaureate nursing education.

Although this work has focused specifically on a pediatric patient interview, the interface could be adapted to suit any in-person interview setting where questions or answers can be scored. Future work could be to apply the principles learned in the design and implementation of SIDNIE and its corresponding scenario builder tool to allow experts in other domains to create their own interpersonal scenarios in virtual reality.

9.1.2 Realism

Due to the format of the SIDNIE system where each question is pre-scored and users are given feedback based on that score, the user is limited to selecting questions that have already been scored. However, as speech recognition and natural language processing advance, using techniques such as machine learning and sentiment analysis, it may be possible in the future to allow users to ask their own questions to the virtual patients and get automated feedback. Another interesting avenue to investigate is whether simply having learners read their selected questions out loud makes a difference in learning outcomes in a scaffolded setting. Our previous work [Pence et al., 2013] shows that it made no difference in a single level's exposure, but future research may show that it makes a difference in learning retention or self efficacy when used in combination with scaffolding.

The goal of this work was to test a scenario's effectiveness for learning apart from any nonverbal cues specific to the situation. Adding in symptom-appropriate animations and providing nurse educators an appropriate interface to determine how they are displayed could be a rich area of future research. In addition to the visual fidelity, both students and professors complained about the low quality of the text to speech voices. We used text-to-speech voices so that new questions and answers could be modified in real time by nurses using the scenario builder tool. Adding the capability to record speech instead would greatly increase workload, but might also yield better learning outcomes or user satisfaction.

9.1.3 Supporting Learning

In this work we tested two scaffolding configurations, both of which were shown to be somewhat effective for learning. Future work may show that an alternate scaffolding method yields improved learning outcomes or user satisfaction.

9.1.4 Use as an Experiment Platform

Because the scenario builder tool allows for quick generation of scenarios with a wide range of characters, it can be used as a platform for the design and execution of many other experiments. Experiments investigating differences in medical interactions with characters with differing demographics such as ethnicity, age, and weight can make immediate use of the scenario builder tool. Even for experiments that require more drastic alterations to character appearances or behaviors, or a change in interaction context, scenarios created using the scenario builder tool can provide a starting point so that a technical expert can make modifications instead of beginning from scratch.

Appendices

Appendix A List of Related Publications

1. Dukes, L. and Hodges, L. F. (2014). Poster: Development of a scenario generation tool for scaffolded virtual patients. In *Proceedings of IEEE Virtual Reality 2014 (IEEE VR 2014)*, pages 131–132

2. Dukes, L. C., Pence, T. B., Hodges, L. F., Meehan, N., and Johnson, A. (2013). Sidnie: scaffolded interviews developed by nurses in education. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 395–406. ACM

3. Pence, T., Dukes, L. C., Hodges, L. F., Meehan, N., and Johnson, A. (2014). An eye tracking evaluation of a virtual pediatric patient training system for nurses. In *Proceedings of Intelligent Virtual Agents 2014*

4. Pence, T. B., Dukes, L. C., Hodges, L. F., Meehan, N. K., and Johnson, A. (2013). The effects of interaction and visual fidelity on learning outcomes for a virtual pediatric patient system. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pages 209–218. IEEE

5. Bloodworth, T., Cairco, L., McClendon, J., Hodges, L. F., Babu, S., Meehan, N. K., Johnson, A., and Ulinski, A. (2011). Initial evaluation of a virtual pediatric patient system. In *Carolinas Women in Computing*

Appendix B Informed Consent: Usability Evaluation

Evaluation of a Virtual Patient Scenario Creation System

Description of the Research and Your Participation You are invited to participate in a research study conducted by Larry F. Hodges, Nancy K. Meehan, Lauren Cairco Dukes, and Toni Bloodworth Pence. The purpose of this research is to evaluate a scenario creation tool for our Pediatric Virtual Patient System. This system would be used to help educate student nurses in evaluation and interview skills when dealing which pediatric patients. There is limited research related to the use of simulation for virtual patients, such as a mother and child. A virtual patient is an artificial intelligent human representation that behaves similarly to an actual human patient under the same set of circumstances, including health care encounters. Pediatrics is a large part of the medical population and it is essential that nurses master the skills necessary to work with children. We believe that this software could be a good method for this purpose and potentially replace or complement the currently used methods.

The researchers will be happy to answer any questions for you. Your participation will involve:

1. You will be assigned a random number that all your responses will be associated with.
2. Your responses and actions will be audio and video recorded.
3. You will answer questions about your background.
4. You will be given a tutorial on how to use the system and what your task will be.
5. You will complete a scenario creation task using the system.
6. The experimenter will administer a questionnaire about the system's usability.
7. You will complete another scenario creation task using the system.
8. The experimenter will administer a questionnaire about the system's usability.
9. You will complete one more scenario creation task.
10. Once completed, you will answer questions about the environment, virtual patients and your interactions.

11. Finally, one of the researchers will interview you to ask you how you felt about the system, your interaction, and any other comments you might have.

The amount of time of your participation will be approximately 1 hour or less.

Potential Benefits The benefits of this research are that you will be able to experience participation in a research study and have the opportunity to interact with virtual characters in a simulated environment. You will also be given the opportunity to be a part of a study that will help contribute to the broader questions of using virtual characters to represent virtual patients in educating nurses in evaluative and interview skills compares to the methods that are currently in use. The results of this research may have an impact on how people use virtual humans for education.

Incentives There are no incentives associated with this study.

Risks and Discomforts There are no known major risks associated with this research. There is a minor risk that you may experience minor discomfort in your eyes, but no more than if you were playing a video game for an hour or watching an hour of television. Resting periods will be provided. If you experience any discomfort, you may discontinue participation at any time without penalty. Another minor risk is that your assigned number may become connected to your responses. To minimize this risk, the associated numbers and consent forms will be kept in a separate locked cabinet for one year and all digitized data will be stored on a password-protected computer, all of which researchers of this study only have access to. After the duration of approval of this study, all physical and digitized data will be destroyed.

Protection of Confidentiality We will do everything we can to protect your privacy. Consent forms and associated numbers will be locked in a cabinet and other coded digitized data will be stored on a password protected computer, all of which researchers of this study only have access to. Any information collected will be summarized across all participants in this research, so that no information will be presented that may identify you specifically. Your identity will not be revealed in any publication that might result from this study. In rare cases, a research study will be evaluated by an oversight agency, such as the Clemson University Institutional Review Board or the federal Office for Human Research Protections, which would require that we share the information we collect

from you. If this happens, the information would only be used to determine if we conducted this study properly and adequately protected your rights as a participant.

Voluntary Participation Your participation in this research study is voluntary. You may choose not to participate and you may withdraw your consent to participate at any time. You will not be penalized in any way should you decide not to participate or to withdraw from this study.

Contact Information If you have any questions or concerns about this study or if any problems arise, please contact Dr. Larry F. Hodges at Clemson University at LFH@clemson.edu or 864.656.7552. If you have any questions or concerns about your rights as a research participant, please contact the Clemson University Office of Research Compliance (ORC) at 864-656-6460 or irb@clemson.edu. If you are outside of the Upstate South Carolina area, please use the ORCs toll-free number, 866-297-3071.

Consent I have read this consent form and have been given the opportunity to ask questions. I give my consent to participate in this study. To participate in this study, I assure that:

- I am at least 18 years of age.
- I have 20/20 vision or corrected to 20/20 vision.
- I use English as my first language and/or am able to communicate in English well.
- I have full use of hearing or corrected hearing with use of a hearing aid in at least one ear.

Participant's signature: _____ Date: _____

A copy of this consent form will be given to you.

Appendix C Demographic Questionnaire: Usability Evaluation

Rate on a scale from 0 (no experience) to 5 (very much experience) your experience level with the following:

- Teaching undergraduate nursing courses
- Interacting with pediatric patients
- Interacting with computers
- Using virtual reality simulations or games
- Developing nursing training simulations

What are the courses you typically teach?

Appendix D Nursing Interviewing Skills Rubric

This material is taken from Diers [Diers, 2008].

D.1 Empathy

Empathy is being aware or attentive to the patient's situation, feelings, concerns, expectations, and/or experiences.

Dimension	Description
Understanding	Words that demonstrate knowledge, awareness, or comprehension of the patient's concerns, expectations, and emotional impact
Legitimize	Words that acknowledge, affirm, or authenticate the patient's feelings, fears, experiences, and/or concerns
Supportive	Words or verbal inquiries that demonstrate hope, comfort, compassion, encouragement, optimism, and/or reassurance to the patient
Sensitive	Words or verbal inquiries that demonstrate an awareness and responsiveness to the needs and feelings of the patient

D.2 Demeanor

Demeanor is the way one behaves or conducts oneself.

Dimension	Description
Open	Words that demonstrate an accessibility, availability, genuineness; also demonstrates candor, honesty, and trust
Cheerful	Words that demonstrate friendly, pleasant, amicable, and polite behavior
Encouraging	Words that are positive, reassuring, and/or inspiring
Confident	Words that express certainty and comfort

D.3 Respect

To respect the patient is to demonstrate courteous regard or value.

Dimension	Description
Consideration	Using words that are understandable, open, patient centered, and not overwhelming
Interest	Words that demonstrate active listening, concern, commitment, and attention
Collaboration	Words that demonstrate an interactive an equal partnership
Empowering	Words that facilitate power, authority, or permission
Relational	Words that demonstrate connection, sharing, and equal exchange

D.4 Credible

To be credible is to be believable, trustworthy, and reliable.

Dimension	Description
Correct	Using spoken words that are free from mistake or error in conveying information (who, what, when, where, why, and how)
Complete	Using spoken words that include all the necessary components, options, evidence, possibilities, and/or steps of the subject matter, focus, or topic at hand
Concise	Using spoken words and phrases that are succinct and brief yet comprehensive when conveying, seeking, or obtaining information
Concrete	Using spoken words and phrases that are specific, objective, informative, and understandable to the receiver when conveying, seeking, or obtaining information

D.5 Coherence

To be coherent is to be logically connected, orderly, and aesthetically consistent.

Dimension	Description
Structure	Using verbal approaches that are logically ordered, use follow-up questions, build on responses, and have a reciprocity of information exchange, avoid blocking behaviors, and avoid overwhelming the patient when conveying, seeking, or obtaining information
Context	Spoken words that assure a continued shared context when conveying, seeking, or obtaining information
Technique	Using advantageous verbal approaches (laundry list, open/closed ended questions, rephrasing, well-placed phrases, inferring, biased or leading questions) when conveying, seeking, or obtaining information

D.6 Clarity

Clear speech is free of confusion and ambiguity and is understandable.

Dimension	Description
Language	Spoken words adapted to the patient's ability to comprehend (right amount, right depth, right language–not too technical, right pace)
Comprehension	Using questions, paraphrasing, and summarizing to check for understanding
Follow Through	Using adaptations to following (repeat, explain more)

Appendix E Postquestionnaire: Participatory Design

Rate on a scale from 0 (no experience) to 5 (very much experience) your experience level with the following:

- Teaching undergraduate nursing courses
- Interacting with pediatric patients
- Interacting with computers
- Using virtual reality simulations or games
- Developing nursing training simulations

Please answer the following questions:

- What are the courses you typically teach?
- What problems did you encounter using this prototype?
- What is missing in this prototype that you think should be added?
- What in this prototype did you want to be able to do, but couldn't?
- How well do you feel like this prototype guided you through completing a scenario?
- What as the most difficult, tedious, or confusing part of using this prototype?
- What was the most fun part of using this prototype?
- What could I add to this design that would make it more useful for nurse educators, or that would make it yield better scenarios?
- What is the highest priority item that needs to change in this prototype?
- Is there anything else about your interaction with the prototype that you want to tell me about?

Please rank the following statements from 1 (completely disagree) to 5 (completely agree):

- I understood what I was supposed to accomplish using this prototype.

- I thought this prototype was easy to use.
- There were a lot of things about this prototype I didn't understand.
- I enjoyed using this prototype.
- Using this prototype was frustrating.

Appendix F Case Study: Usability Evaluation

Copied from `http://www.hawaii.edu/medicine/pediatrics/pedtext/s01c01.html`. Accessed 2/9/2015.

A six year old male presents to your office for his annual well child visit. He is accompanied by his mother. You have cared for this child since his birth, and he has had regular well child care. You last saw him for his visit prior to entering kindergarten at age five years. Today his mother notes that she has been anxiously awaiting this visit as she has several concerns to discuss:

He is having some difficulty in school (now just finishing the first quarter of first grade). He is struggling to learn to read, and has some difficulty with arithmetic. His teacher called his mother yesterday to report that he hasn't been turning in his homework or completing his classroom assignments. His mother indicates that she was very surprised to hear this, as the previous teacher reports have indicated that he was doing adequate work.

He has frequent complaints of stomachache. He has a good appetite, but has always been a "picky eater". He enjoys drinking milk.

He has been having increasing nasal congestion over the last few months. He has had some sneezing attacks, and seems to clear his throat often. He does cough at night. The cough often sounds "wet" to his mother. He also joins in to tell you that he has a hard time breathing during PE. He has no other regular physical activity, but his mother reports that he is always "busy doing something". His mother reminds you that he was born prematurely at 34 weeks, and had difficulty with wheezing as a younger child, but he has done well in the last year or two and hasn't needed any medications for wheezing.

Exam: VS are normal. Weight 30 kg (¿ 95%ile), height 117 cm (46") (50%). In general, he appears to be an overweight, friendly child who is cooperative and who appears to be his stated age. He is active in the exam room, exploring the contents of the drawers and cabinets. He interrupts his mother repeatedly during the interview. He appears to be mouth breathing with significant nasal congestion. His tonsils are large but not inflamed. His heart is regular without murmurs. His pulses are normal. His lungs have clear breath sounds, with transmitted upper airway rhonchi. There are no wheezes, but the I:E ratio is prolonged. His abdominal and neurologic screening exams are normal.

Appendix G Post-Task Questionnaire: Usability Evaluation

After each of the three tasks, the nurse will answer these two questions: On a scale of 0 (very easy) to 5 (very difficult),

- How difficult was the task you just completed?
- Do you have any suggestions for how the system could be improved to better facilitate the task?

Appendix H Post-Interview: Usability Evaluation

Rate on a scale from 0 (no experience) to 5 (very much experience) your experience level with the following:

- Which of the three tasks you were assigned was the easiest? Why?
- Which of the three tasks you were assigned was the hardest? Why?
- What did you like the best about the system you used?
- What did you dislike the most about the system you used?
- If you could change it, how would you change the system?
- Is there anything else you would like us to know about your interaction with the system?

Appendix I System Usability Scale

This questionnaire is adapted from the System Usability Scale [Brooke, 1996].

After all three tasks are completed, the nurses will answer the following questions: Please answer the following about this system, scoring from 0 (completely disagree) to 5 (completely agree).

Please rank the following questions on a scale of 1 (strongly disagree) to 5 (strongly agree).

- I think that I would like to use this system frequently.
- I found the system unnecessarily complex.
- I thought the system was easy to use.
- I think that I would need the support of a technical person to be able to use this system.
- I found the various functions in this system were well integrated.
- I thought there was too much inconsistency in this system.
- I would imagine that most people would learn to use this system very quickly.
- I found the system very cumbersome to use.
- I felt very confident using the system.
- I needed to learn a lot of things before I could get going with this system.

Appendix J Informed Consent: Scaffolding Evaluation

Evaluating The Impact of Scaffolded Presentation of Pediatric Interviewing Scenarios

Description of the Research and Your Participation You are invited to participate in a research study conducted by Larry F. Hodges, Nancy K. Meehan, Lauren Cairco Dukes, and Toni Bloodworth Pence. The purpose of this research is to evaluate a scenario creation tool for our Pediatric Virtual Patient System. This system would be used to help educate student nurses in evaluation and interview skills when dealing which pediatric patients. There is limited research related to the use of simulation for virtual patients, such as a mother and child. A virtual patient is an artificial intelligent human representation that behaves similarly to an actual human patient under the same set of circumstances, including health care encounters. Pediatrics is a large part of the medical population and it is essential that nurses master the skills necessary to work with children. We believe that this software could be a good method for this purpose and potentially replace or complement the currently used methods.

The researchers will be happy to answer any questions for you. Your participation will involve:

1. You will be assigned a random number that all your responses will be associated with.
2. Your responses and actions will be audio and video recorded.
3. You will answer questions about your background.
4. You will take a pre-test on your interviewing skills.
5. You will be given a tutorial on how to use the system and what your task will be.
6. After you have been instructed, you will be permitted to interact with the system until you have come up with a conclusion or the allotted time has expired.
7. Once completed, you will complete a post-test of your interviewing skills, and will then answer questions about the environment, virtual patients and your interactions.
8. Finally, one of the researchers will interview you to ask you how you felt about the system, your interaction, and any other comments you might have.

The amount of time of your participation will be approximately 1 hour or less.

Potential Benefits The benefits of this research are that you will be able to experience participation in a research study and have the opportunity to interact with virtual characters in a simulated environment. You will also be given the opportunity to be a part of a study that will help contribute to the broader questions of using virtual characters to represent virtual patients in educating nurses in evaluative and interview skills compares to the methods that are currently in use. The results of this research may have an impact on how people use virtual humans for education.

Incentives You may be offered extra credit for your participation in this study. It is the professor discretion to decide if extra credit will be offered and how much the extra credit is worth. If extra credit is offered by a professor for participation in this study, there will also be an equivalent alternative offered for those students who do not wish to participate.

Risks and Discomforts There are no known major risks associated with this research. There is a minor risk that you may experience minor discomfort in your eyes, but no more than if you were playing a video game for an hour or watching an hour of television. Resting periods will be provided. If you experience any discomfort, you may discontinue participation at any time without penalty. Another minor risk is that your assigned number may become connected to your responses. To minimize this risk, the associated numbers and consent forms will be kept in a separate locked cabinet for one year and all digitized data will be stored on a password-protected computer, all of which researchers of this study only have access to. After the duration of approval of this study, all physical and digitized data will be destroyed.

Protection of Confidentiality We will do everything we can to protect your privacy. Consent forms and associated numbers will be locked in a cabinet and other coded digitized data will be stored on a password protected computer, all of which researchers of this study only have access to. Any information collected will be summarized across all participants in this research, so that no information will be presented that may identify you specifically. Your identity will not be revealed in any publication that might result from this study. In rare cases, a research study will be evaluated by an oversight agency, such as the Clemson University Institutional Review Board or the federal Office for Human Research Protections, which would require that we share the information we collect from you. If this happens, the information would only be used to determine if we conducted this study properly and adequately protected your rights as a participant.

Voluntary Participation Your participation in this research study is voluntary. You may choose not to participate and you may withdraw your consent to participate at any time. You will not be penalized in any way should you decide not to participate or to withdraw from this study.

Contact Information If you have any questions or concerns about this study or if any problems arise, please contact Dr. Larry F. Hodges at Clemson University at LFH@clemson.edu or 864.656.7552. If you have any questions or concerns about your rights as a research participant, please contact the Clemson University Office of Research Compliance (ORC) at 864-656-6460 or irb@clemson.edu. If you are outside of the Upstate South Carolina area, please use the ORCs toll-free number, 866-297-3071.

Consent I have read this consent form and have been given the opportunity to ask questions. I give my consent to participate in this study. To participate in this study, I assure that:

- I am at least 18 years of age.
- I have 20/20 vision or corrected to 20/20 vision.
- I use English as my first language and/or am able to communicate in English well.
- I have full use of hearing or corrected hearing with use of a hearing aid in at least one ear.

Participant's signature: _____ Date: _____

A copy of this consent form will be given to you.

Appendix K Demographic Questionnaire: Scaffolding Evaluation

Please fill out the following information about yourself:

- Age
- Gender
- Occupational Status
- Ethnicity

Rate on a scale from 0 (no experience) to 7 (very much experience) your experience level with the following:

- Health care experience
- Experience working with children
- Experience working in a health care setting
- Experience working with children in a health care setting
- Using a computer in your daily activities
- Exposure to virtual humans, intelligent agents, or avatars in general
- Using virtual reality or simulated environments in your daily activities

Please briefly describe in what way have you been exposed to or used virtual humans, intelligent agents, or avatars in general.

Please list any previous nursing courses you have taken.

Appendix L New General Self-Efficacy Scale

This scale is taken from Chen et al. [Chen et al., 2001].

Please respond to the following questions rating them from 1 (strongly disagree) to 5 (strongly agree).

- I will be able to achieve most of the goals I have set for myself.
- When facing difficult tasks, I am certain that I will accomplish them.
- In general, I think I can obtain outcomes that are important to me.
- I believe I can succeed at most any endeavor to which I set my mind.
- I will be able to successfully overcome many challenges.
- I am confident that I can perform effectively on many different tasks.
- Compared to other people, I can do most tasks very well.
- Even when things are tough, I can perform quite well.

Appendix M Learning Outcomes: Scaffolding Evaluation

Your patient today is a four year old boy named Patrick Jones. He is in the doctor's office with his mother, Aliceann Jones. Patrick's electronic health record states that the reason for today's visit is that he has had stomach pain for the past two days.

Please write ten questions that you would ask Patrick that meet these two criteria:

- sensitive, which means that it uses words that demonstrate knowledge, awareness, or comprehension of the patient's concerns, expectations, and emotional impact
- appropriate language, which means that the spoken words are adapted to the patient's ability to comprehend (right amount, right depth, right language–not too technical, right pace)

Please score these ten questions on those two criteria:

- Hello, my name is (your name). Can you tell me your names?
- Patrick, how frequent is your stomach pain?
- There's a really terrible virus going around. Please don't tell me Patrick's been
- vomiting.
- Patrick, can you rate your stomach pain on a scale of 0 to 10, where 0 is no pain and 10 is like you are being eaten by a shark?
- Patrick, is there anything you do that makes your tummy feel better?
- Mrs. Jones, has Patrick exhibited any other gastrointestinal or genitourinary symptoms?
- Mrs. Jones, you're not one of those anti vaccine parents, are you?
- Patrick, are you taking any medications?
- Mrs. Jones, is Patrick allergic to any medications?
- Patrick, can you point to where your stomach hurts?

Do you feel like you used a different style of speaking when writing questions for Patrick than you did when interacting with SIDNIE?

Appendix N Presence and Copresence Questionnaires

This questionnaire is adapted from the Slater co-presence questionnaire found in [Mortensen et al., 2002].

N.1 Presence

Please rank the following questions on a scale of 1 (not at all) to 7 (a great deal).

- To what extent did you feel like you were actually in the room with the virtual patients?
- How dizzy, sick or nauseous did you feel resulting from the experience, if at all?
- To what extent were there times during the experience when the virtual room was reality for you?

Please rank the following question from 1 (images that I saw) to 7 (somewhere that I visited).

- When you think back about your experience, do you think of the virtual room more as images that you saw, or more as somewhere that you visited?

N.2 Copresence

Please rank the following questions on a scale of 1 (not at all) to 7 (a great deal).

- 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?
- To what extent, if at all, did the virtual patients remind you of someone you've met in the real world?
- To what extent, if at all, did the virtual patients hinder you from carrying out the task?
- To what extent, if at all, did the virtual patients help you in carrying out the task?
- To what extent, if at all, did you have a sense of working with the virtual patients?

- 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?
- To what extent, if at all, did you feel that the virtual patients were listening to you?
- How realistic were the virtual patients (for example, how they looked, moved, spoke and interacted with you)?
- To what extent, if at all, were you worried about what the virtual patients thought of your performance?

Please rank the following questions on a scale of 1 (strongly disagree) to 7 (strongly agree).

- I felt that the virtual patients were watching me.
- The thought that the virtual patients were not real people crossed my mind often.
- The virtual patients appeared to be sentient (conscious and alive) to me.
- I perceived the virtual patients as only computerized images, not as people.

Appendix O Post-Interview: Scaffolding Evaluation

- What did you think about this system?
- How well do you think you learned from this system?
- What did you like the best about interacting with SIDNIE?
- What did you dislike the most about interacting with SIDNIE?
- Is there anything we could have done better to help you learn?
- Did you feel like the patients were realistic?
- Did you feel like the scenarios were realistic?
- Do you think you would use this system again for practice?
- Do you think using this system will affect your clinical practice?

Bibliography

[Anderson et al., 2005] Anderson, L. W., Krathwohl, D. R., and Bloom, B. S. (2005). *A taxonomy for learning, teaching, and assessing*. Longman.

[Ball et al., 2013] Ball, J. W., Bindler, R. C., and Cowen, K. J. (2013). *Child health nursing*. Prentice Hall.

[Banerjee and Pedersen, 2002] Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pages 136–145. Springer.

[Benedict, 2010] Benedict, N. (2010). Virtual patients and problem-based learning in advanced therapeutics. *American journal of pharmaceutical education*, 74(8).

[Bird, 2006] Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

[Blender, 2012] Blender (2012). Blender. `http://www.blender.org`.

[Bloodworth et al., 2011] Bloodworth, T., Cairco, L., McClendon, J., Hodges, L. F., Babu, S., Meehan, N. K., Johnson, A., and Ulinski, A. (2011). Initial evaluation of a virtual pediatric patient system. In *Carolinas Women in Computing*.

[Booth and McMullen-Fix, 2012] Booth, T. L. and McMullen-Fix, K. (2012). Innovation center: Collaborative interprofessional simulation in a baccalaureate nursing education program. *Nursing Education Perspectives*, 33(2):127–129.

[Botezatu et al., 2010] Botezatu, M., Hult, H., Tessma, M. K., and Fors, U. G. (2010). Virtual patient simulation for learning and assessment: Superior results in comparison with regular course exams. *Medical Teacher*, 32(10):845–850.

[Bray et al., 2011] Bray, B. S., Schwartz, C. R., Odegard, P. S., Hammer, D. P., and Seybert, A. L. (2011). Assessment of human patient simulation-based learning. *American journal of pharmaceutical education*, 75(10).

[Brooke, 1996] Brooke, J. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189:194.

[Chen et al., 2001] Chen, G., Gully, S. M., and Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods*, 4(1):62–83.

[Consorti et al., 2012] Consorti, F., Mancuso, R., Nocioni, M., and Piccolo, A. (2012). Efficacy of virtual patients in medical education: A meta-analysis of randomized studies. *Computers & Education*, 59(3):1001–1008.

[Cook and Triola, 2009] Cook, D. A. and Triola, M. M. (2009). Virtual patients: a critical literature review and proposed next steps. *Medical education*, 43(4):303–311.

[Dao and Simpson, 2005] Dao, T. N. and Simpson, T. (2005). Measuring similarity between sentences. `http://opensvn.csie.org/WordNetDotNet/trunk/Projects/Thanh/Paper/WordNetDotNet_Semantic_Similarity.pdf`.

[Deladisma et al., 2009] Deladisma, A. M., Gupta, M., Kotranza, A., Bittner IV, J. G., Imam, T., Swinson, D., Gucwa, A., Nesbit, R., Lok, B., Pugh, C., et al. (2009). A pilot study to integrate an immersive virtual patient with a breast complaint and breast examination simulator into a surgery clerkship. *The American Journal of Surgery*, 197(1):102–106.

[DI-Guy, 2009] DI-Guy (2009). Di-guy human simulation software. `http://www.diguy.com`.

[Diers, 2008] Diers, J. E. (2008). *Assessing and appraising nursing students' professional communication*. ProQuest Dissertations.

[Dukes and Hodges, 2014] Dukes, L. and Hodges, L. F. (2014). Poster: Development of a scenario generation tool for scaffolded virtual patients. In *Proceedings of IEEE Virtual Reality 2014 (IEEE VR 2014)*, pages 131–132.

[Dukes et al., 2013] Dukes, L. C., Pence, T. B., Hodges, L. F., Meehan, N., and Johnson, A. (2013). Sidnie: scaffolded interviews developed by nurses in education. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 395–406. ACM.

[Durham and Alden, 2008] Durham, C. F. and Alden, K. R. (2008). Enhancing patient safety in nursing education through patient simulation. *Patient safety and quality: An evidence-based handbook for nurses*, 6(3):221–250.

[Ellaway, 2010] Ellaway, R. H. (2010). Openlabyrinth: An abstract pathway-based serious game engine for professional education. In *Digital Information Management (ICDIM), 2010 Fifth International Conference on*, pages 490–495. IEEE.

[EMRSoap, 2014] EMRSoap (2014). Soap notes. `http://www.emrsoap.com/definitions/soap/`.

[Faulkner, 2003] Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3):379–383.

[Forehand, 2010] Forehand, M. (2010). Blooms taxonomy. *Emerging perspectives on learning, teaching, and technology*.

[Forsberg et al., 2011] Forsberg, E., Georg, C., Ziegert, K., and Fors, U. (2011). Virtual patients for assessment of clinical reasoning in nursinga pilot study. *Nurse Education Today*, 31(8):757–762.

[Hadden, 2009] Hadden, D. (2009). An expert systems-based virtual patient simulation system for assessing and mentoring clinician decision making: Acceptance, reach and outcomes. *The Journal for Simulation in Healthcare*, 4:238–239.

[Harder, 2010] Harder, B. N. (2010). Use of simulation in teaching and learning in health sciences: A systematic review. *The Journal of nursing education*, 49(1):23.

[Hockenberry et al., 2007] Hockenberry, M. J., Wilson, D., et al. (2007). *Wong's nursing care of infants and children*. Mosby/Elsevier.

[Holton and Clarke, 2006] Holton, D. and Clarke, D. (2006). Scaffolding and metacognition. *International journal of mathematical education in science and technology*, 37(2):127–143.

[Hubal et al., 2003] Hubal, R. C., Deterding, R. R., Frank, G. A., Schwetzke, H. F., and Kizakevich, P. N. (2003). Lessons learned in modeling virtual pediatric patients. *Studies in Health Technology and Informatics*, pages 127–130.

[IVONA, 2012] IVONA (2012). Ivona text-to-speech voices — tts voices — text to voice. `http://www.ivona.com/us/`.

[Johnsen et al., 2005] Johnsen, K., Dickerson, R., Raij, A., Lok, B., Jackson, J., Shin, M., Hernandez, J., Stevens, A., and Lind, D. S. (2005). Experiences in using immersive virtual characters to educate medical communication skills. In *Virtual Reality, 2005. Proceedings. VR 2005. IEEE*, pages 179–186. IEEE.

[Johnsen et al., 2007] Johnsen, K., Raij, A., Stevens, A., Lind, D. S., and Lok, B. (2007). The validity of a virtual human experience for interpersonal skills education. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1049–1058. ACM.

[Kujala, 2003] Kujala, S. (2003). User involvement: a review of the benefits and challenges. *Behaviour & information technology*, 22(1):1–16.

[Lakens, 2013] Lakens, D. (2013). Calculating and reporting effet sizes to facilitate cumulative science: A practical primer for t-tests and anovas.

[Lok, 2006] Lok, B. (2006). Teaching communication skills with virtual humans. *Computer Graphics and Applications, IEEE*, 26(3):10–13.

[Lua et al., 2009] Lua, J., LaJoie, S. P., and Wiseman, J. (2009). Scaffolding collaboration in simulated medical emergencies. In *Anonymous International Conference on Computers in Education.(). Hong Kong: Asia-Pacific Society for Computers in Education*, pages 285–292.

[Magnusson, 2014] Magnusson, K. (2014). Interpreting cohen's d effect size: an interactive visualization. `http://rpsychologist.com/d3/cohend/`.

[MakeHuman, 2013] MakeHuman (2013). Makehuman open source tool for making 3d characters. `http://makehuman.org/`.

[Medbiquitous, 2007] Medbiquitous (2007). Medbiquitous working group virtual patient summary. `http://www.medbiq.org/working_groups/virtual_patient/MedBiquitousVirtualPatientSummary.pdf`.

[Melzack and Katz, 2007] Melzack, R. and Katz, J. (2007). Mcgill pain questionnaire. In *Encyclopedia of Pain*, pages 1102–1104. Springer.

[Miller, 1995] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

[Mortensen et al., 2002] Mortensen, J., Vinayagamoorthy, V., Slater, M., Steed, A., Lok, B., and Whitton, M. (2002). Collaboration in tele-immersive environments. In *Proceedings of the workshop on Virtual environments 2002*, pages 93–101. Eurographics Association.

[Nielsen and Landauer, 1993] Nielsen, J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 206–213. ACM.

[of Health et al., 2014] of Health, D., for Medicare, H. S. C., and Services, M. (2014). Evaluation and management services guide. http://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/eval_mgmt_serv_guide-ICN006764.pdf.

[Pence et al., 2014] Pence, T., Dukes, L. C., Hodges, L. F., Meehan, N., and Johnson, A. (2014). An eye tracking evaluation of a virtual pediatric patient training system for nurses. In *Proceedings of Intelligent Virtual Agents 2014*.

[Pence et al., 2013] Pence, T. B., Dukes, L. C., Hodges, L. F., Meehan, N. K., and Johnson, A. (2013). The effects of interaction and visual fidelity on learning outcomes for a virtual pediatric patient system. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pages 209–218. IEEE.

[Planas and Nelson, 2008] Planas, L. G. and Nelson, L. E. (2008). A systems approach to scaffold communication skills development. *American journal of pharmaceutical education*, 72(2).

[Posel et al., 2009] Posel, N., Fleiszer, D., and Shore, B. M. (2009). 12 tips: Guidelines for authoring virtual patient cases. *Medical Teacher*, 31(8):701–708.

[Poser, 2012] Poser (2012). Poser 3d animation & character creation software - official website. http://poser.smithmicro.com/.

[Quintana and Fishman, 2006] Quintana, C. and Fishman, B. (2006). Supporting science learning and teaching with software-based scaffolding. *American Educational Research Association (AERA)*.

[Rajak and Saxena, 2010] Rajak, A. and Saxena, K. (2010). Achieving realistic and interactive clinical simulation using case based therasims therapy engine dynamically. In *Proceedings of the National Conference on Advanced Pattern Mining and Multimedia Computing*, pages 624–628.

[Rauterberg et al., 1995] Rauterberg, M., Strohm, O., and Kirsch, C. (1995). Benefits of user-oriented software development based on an iterative cyclic process model for simultaneous engineering. *International Journal of Industrial Ergonomics*, 16(4):391–409.

[Reese et al., 2010] Reese, C. E., Jeffries, P. R., and Engum, S. A. (2010). Learning together: Using simulations to develop nursing and medical student collaboration. *Nursing education perspectives*, 31(1):33–37.

[Rezzable, 2012] Rezzable (2012). Metamorph for unity — rezzable. http://rezzable.com/metamorph/manuals/unity.

[Rossen et al., 2009] Rossen, B., Lind, S., and Lok, B. (2009). Human-centered distributed conversational modeling: Efficient modeling of robust virtual human conversations. In *Intelligent Virtual Agents*, pages 474–481. Springer.

[Rossen and Lok, 2012] Rossen, B. and Lok, B. (2012). A crowdsourcing method to develop virtual human conversational agents. *International Journal of Human-Computer Studies*, 70(4):301–319.

[Round et al., 2009] Round, J., Conradi, E., and Poulton, T. (2009). Training staff to create simple interactive virtual patients: the impact on a medical and healthcare institution. *Medical Teacher*, 31(8):764–769.

[Sakpal, 2012] Sakpal, R. (2012). Virtual patients to teach cultural competency. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 349–352. ACM.

[Sanders and Welk, 2005] Sanders, D. and Welk, D. S. (2005). Strategies to scaffold student learning: Applying vygotsky's zone of proximal development. *Nurse Educator*, 30(5):203–207.

[SAPI, 2013] SAPI (2013). Microsoft speech api (sapi) 5.3. `http://msdn.microsoft.com/en-us/library/ms723627(v=vs.85).aspx`.

[Sawyer, 2006] Sawyer, R. K. (2006). *The Cambridge handbook of the learning sciences*, volume 2:5. Cambridge University Press New York.

[Sharp et al., 2007] Sharp, H., Rogers, Y., and Preece, J. (2007). Interaction design: beyond human-computer interaction. *West Sussex, England: John Wiley & Sons*.

[SIMmersion, 2013] SIMmersion (2013). Simmersion, llc immersive simulations. `http://www.simmersion.com`.

[Software, 2015] Software, F. F. (2015). Modern ui for wpf. `https://github.com/firstfloorsoftware/mui`.

[Spouse, 1998] Spouse, J. (1998). Scaffolding student learning in clinical practice. *Nurse Education Today*, 18(4):259–266.

[SQLite, 2012] SQLite (2012). Sqlite home page. `http://www.sqlite.org/`.

[Stevens et al., 2006] Stevens, A., Hernandez, J., Johnsen, K., Dickerson, R., Raij, A., Harrison, C., DiPietro, M., Allen, B., Ferdig, R., Foti, S., et al. (2006). The use of virtual patients to teach medical students history taking and communication skills. *The American Journal of Surgery*, 191(6):806–811.

[Stupans and Owen, 2009] Stupans, I. and Owen, S. (2009). *Planning and scaffolding for learning in experiential placements in Australian pharmacy schools*. PhD thesis, New Zealand Association for Coop Education.

[Thalmann, 2001] Thalmann, D. (2001). The role of virtual humans in virtual environment technology and interfaces. In *Frontiers of Human-Centered Computing, Online Communities and Virtual Environments*, pages 27–38. Springer.

[Tilley et al., 2007] Tilley, D. S., Allen, P., Collins, C., Bridges, R. A., Francis, P., and Green, A. (2007). Promoting clinical competence: Using scaffolded instruction for practice-based learning. *Journal of Professional Nursing*, 23(5):285–289.

[Trepagnier et al., 2011] Trepagnier, C. Y., Olsen, D. E., Boteler, L., and Bell, C. A. (2011). Virtual conversation partner for adults with autism. *Cyberpsychology, Behavior, and Social Networking*, 14(1-2):21–27.

[Unity3D, 2012] Unity3D (2012). Unity: Game development tool. http://www.unity3d.com.

[Wilson, 2006] Wilson, D. M. (2006). Itech: an interactive technical assistant. *ProQuest Dissertations*.