Usability Evaluation of a Pediatric Virtual Patient Creation Tool

Lauren Cairco Dukes
School of Computing
Clemson University
Clemson, SC 29634
Email: LCairco@clemson.edu

Nancy Meehan
School of Nursing
Clemson University
Clemson, SC 29634
Email: nmeehan@clemson.edu

Larry F. Hodges
School of Computing
Clemson University
Clemson, SC 29634
Email: LFH@clemson.edu

***Abstract*—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. Previously, we worked with nurse educators in a participatory design process to create a scenario builder tool for SIDNIE [paper in submission]. In this work we detail a usability evaluation of the scenario creation tool. We found that nurse educators were able to use our tool to create a virtual patient scenario in less than two hours.**

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. 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 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. Although text-and-photo based virtual patients are easier to produce because they can use standardized images and case studies, much of human communication is nonverbal [24] 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 initial production could be difficult. 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 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 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 cannot provide in a timely manner.

In collaboration with the School of Nursing at Clemson University, 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. SIDNIE's first scenario was effective for learning [6, 18]. 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 independently create new scenarios for SIDNIE. To further measure the end-product's usability, in this paper we detail a usability study of the scenario builder tool. We found that nurses could use the scenario builder tool to create a scenario in less than two hours, and that much of that development time could be outsourced to individuals with fewer skills, such as an undergraduate student.

I. Related Work

Simulation learning that mimics real world scenarios is beneficial to nursing students and provides standardized ex-

978-1-5090-6117-4/16 $31.00 © 2016 IEEE
DOI 10.1109/ICHI.2016.20

periences in which students can practice problem solving techniques and clinical decision making abilities [12]. Virtual patients can be effective for simulation training since they 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) [12]. Medical students have used virtual patient scenarios to develop their communication skills [14, 4], interdisciplinary collaboration [2] and patient interviewing skills [13]. Johnsen et al. [13] 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. 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 [7], where they were used for training and assessment for medical students. Even with the lack of realistic characters, the results of their study were positive in that many of the participants stated that they gained valuable experience.

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

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, nonlinear, or branching), 4) the inclusion of assessment and feedback, and 5) the methods for interactivity [19]. Despite nurses' best efforts, Round et al. [23] noted that difficulties in virtual patient development include inflexibility, non-realistic scenarios, and lack of engaging content in the final products.

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 [22]. This process can be time-consuming. Indeed, implementing the SIDNIE system took nine months, three months beyond the approximate six months or 200 hours to develop a conversationally accurate virtual patient system [22]. 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 [20, 11]. However, their business model does not allow nursing educators to directly develop or change their own scenarios.

A. Other 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 then translates it into simulation data [22]. 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. Round et al. [23] developed an inexpensive method to create multipath virtual patients which were composed of text and photo elements. Similarly, the Decision Simulation [1] and OpenLabyrinth [8] authorship tools provide flexibility, but only image-and-text interaction.

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 a model of "Human-centered Distributed Conversational Modeling" [22], a modified crowdsourcing technique. Unlike conventional design methods in which six months and 200 hours are needed to develop a conversational model that is 75% accurate, the tool can reduce development time to as little as 15 hours [21].

III. System Description

During the participatory design process, we worked with three nurse educators to make a scenario creation tool [paper in submission]. In general, the tool followed the format of a typical medical interview citemedicare with a few additions for the virtual patient character generation. The application roughly followed a "wizard" format, although nurse educators had the freedom to move freely between steps instead of only moving forward or back one step.

The first task was to define the scenario's title and learning goals. Nurses typed a scenario title and a learning goal, then selected criteria that the questions could be scored on from a bank of scoring criteria. Next, the nurse educator selected virtual characters to represent the pediatric patient and accompanying adult. Users are first encouraged to select a character 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 next screen prompted the nurse educator to select a chief complaint from a bank of chief complaints. After that, the nurse educator filled out the history of present illness, which included characteristics of the problem, including timing, pain level, and alleviating factors. After the history of present illness, the nurse educator filled out a review of systems, marking symptoms as present or absent. Using the chief complaint information, the review of systems was already "prefilled" to

show the most common symptoms that occurred with the given chief complaint, but the nurse educator could add or remove symptoms as he or she wished. Finally, 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. The nurse educator could add vital signs as well as upload images or notes.

Using the information the nurse has provided in the preceding steps, the system would then 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. If the nurse educator 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 is contributed to the library.

Not all of the usability concerns of the participants were addressed by the end of the design process, so before our usability study we modified the system according to their concerns. The original design had integrated the tasks of contributing to the database for future reuse and creating new scenarios. However, participants suggested the removal of the nonessential "content contribution" tasks from the flow of the scenario builder tool, as they found them distracting and confusing. We implemented new interfaces for creating chief complaints and creating new scoring criteria, which were notably the most tedious aspects of the creation process. One participant had noted that those tasks could also be outsourced to a less skilled worker, so it made sense to separate those interfaces out. 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.

A. Finding, Creating, and Customizing Characters

In the previous study, when creating characters, participants 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 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

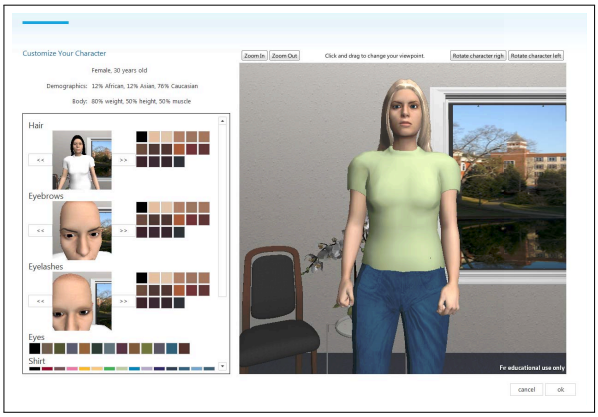

Fig. 1. The updated customization interface shows thumbnail views of hair, eyebrows, and eyelashes with buttons to right and left to change their styles.

rotate their character clockwise or counterclockwise to see the character's profile or back.

To increase the character realism in response to complaints in previous studies, 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 asymmetries from character to character to provide a greater variety of characters.

B. Clarifying Question Previews through SIDNIE Integration

For two out of the three participants in the design study, the question previewer interface was confusing. It showed a list of questions with checkboxes that indicated the question's scoring. 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-

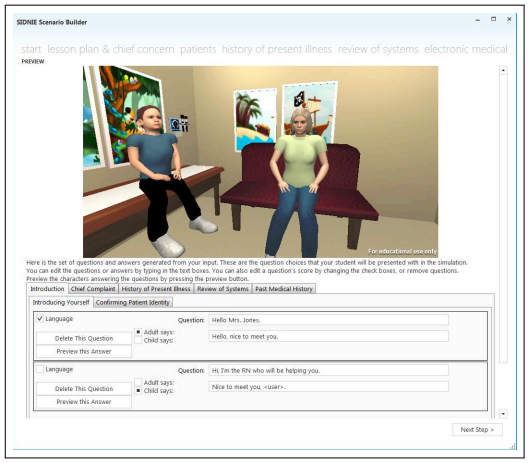

Fig. 2. The updated previewer shows the selected patients in a doctor's office along with questions and answers. When the user clicks the button to preview a question, the patients respond with text-to-speech and lipsyncing.

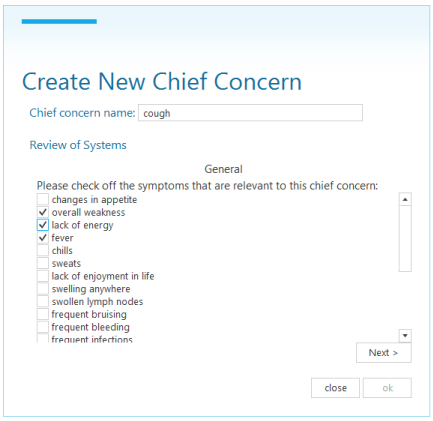

Fig. 3. To create a new chief complaint, for each system, the user checks off the symptoms they consider to be relevant.

to-speech with lip-syncing animations. See Figure 2 for the updated previewer layout.

C. Contributing a New Chief Complaint

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 deems relevant to the chief complaint (Figure 3)–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 4).

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

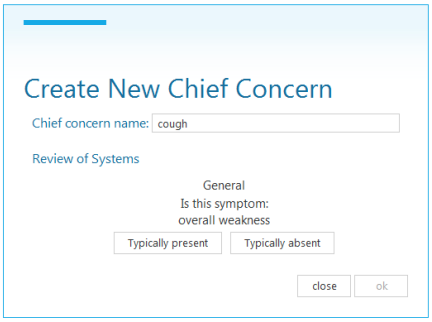

Fig. 4. After selecting the relevant symptoms, the user selects whether each relevant symptom is typically present or typically absent for their chief complaint.

chief complaint is added to the database of available chief complaints.

D. 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 5 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.

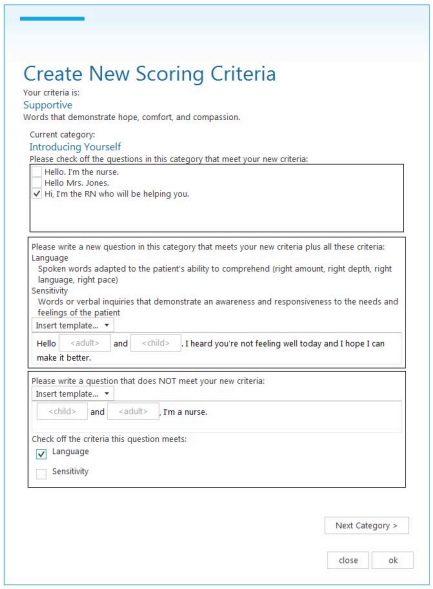

Fig. 5. When creating a new criteria, for each interview category the user is required to score all the existing questions as well as to write two new questions to contribute to the database.

IV. 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 [10, 16]. 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, the participant answered questions about their nursing experiences and experiences with children. We 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. 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 [5]. 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 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[5]. 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 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 [3] and a post-interview to gather information on how participating nursing faculty perceived the system.

Our hypotheses for this experiment were that (1) the participants would find the system easy to use and (2) participants would be able to make a scenario in less than 30 minutes.

V. Results

Seven participants took part in the experiment–one student, and six faculty members. Participants varied widely in every demographic measure (Table I). Two participants participated in the design of the scenario builder tool.

TABLE I
PARTICIPANT DEMOGRAPHIC INFORMATION. IN THE "ROLE" ROW, F STANDS FOR FACULTY, 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 undergraduates	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 simulations	5	2	5	4	1	2	2

A. 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. Table II 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.

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 III for the recorded completion times.

The student participant completed an alternate set of three tasks. 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 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.

B. 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 IV 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).

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 V for the overall summary rankings. The student participant found creating a new chief complaint least difficult and creating a new scoring criteria most difficult.

C. 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 VI 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.

D. Usability Observations

Informally, participants gave feedback throughout their interaction with the system. Participants tended to have the same usability problems throughout the task interaction.

Character Selection and Creation Every participant had difficulty in selecting 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 6). As filters are

TABLE II
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

TABLE III
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

TABLE IV
DIFFICULTY RANKINGS FOR EACH TASK COMPLETED BY FACULTY PARTICIPANTS (1=LEAST DIFFICULT, 5=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

TABLE V
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

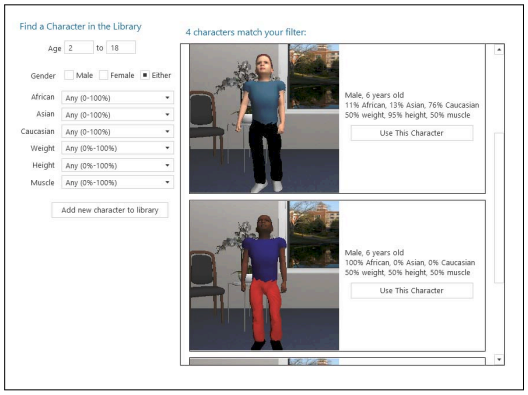

Fig. 6. The user can select a patient from the library or choose to create their own character.

modified, the contents of the box on the right side of the screen are immediately filtered. There were two distinct usability flaws here. First, at least three out of the six faculty participants did not understand that the library was immediately filtered, 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. We could possibly improve usability by adding a search button and disabling 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 character creation dialogue box. The dialogue is instead populated with average values. We designed it this way because most of the filters are presented as a range, while character creation requires specific values. 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.

We observed that all participants, when creating a character, instead of filling out the dialogue in order of its fields, filled out the gender, ethnicity, then age, and finally the body characteristics. This suggests that the controls should possibly be rearranged on the page. 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." Two participants said that they would not think about a patient's muscle percentage in this context. 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.

TABLE VI
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 I 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

When customizing the character, the primary difficulty was in the interface for selecting hair, eyebrows, and eyelashes (Figure 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.

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.

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. 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.

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.

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.

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.

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, or using the default information (Figure 5). 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.

E. Subjective Feedback

Participants also answered usability questions in a post-task 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. One participant suggested a different ordering for the interview, as a SOAP note [9] instead of the Medicare standard [17]. 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 preexisting characters or chief complaints.

VI. Discussion

Our evaluation is limited by the number of participants in the study, especially considering the differences in demographic characteristics. However, our preliminary results are promising. Our 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 indicates that the system was usable overall.

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, 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 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 previewing the scenario might be better split apart from the rest of the task and saved for later review.

Finally, creating a new scoring criteria was a challenge for every user. Besides the interface difficulties, 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 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 needed. Since the student participant was capable of completing each of these tasks, it is possible to reduce that workload by outsourcing those tasks. One participant said, "Time is a big factor when it comes to our work...creating the avatar is fun...let an undergraduate do the nitty gritty and then let the faculty member sign off at the end."

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 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.

VII. Recommendations for Future 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 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. In the future, it seems that the "offline" workload of creating content should be considered a part of the creation task. Providing supplemental resources (for example, worksheets or instructions) could potentially speed along content creation and reduce difficulty.

Another observation 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. 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. 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. In any case, future designers would do well to anticipate "experimentation" and should expect that 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 during the participatory design phase, 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. In this 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 originally participated the participatory design) 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. 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.

Acknowledgements

This work was supported by a NSF Graduate Research Fellowship (fellow ID: 2009080400), 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). The authors also thank Toni Pence for her work on the SIDNIE system.

REFERENCES

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

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

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

[4] A. M. Deladisma, M. Gupta, A. Kotranza, J. G. Bittner IV, T. Imam, D. Swinson, A. Gucwa, R. Nesbit, B. Lok, C. Pugh, et al. 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, 2009.

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

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

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

[8] R. H. Ellaway. 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, 2010.

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

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

[11] D. Hadden. 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, 2009.

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

[13] K. Johnsen, A. Raij, A. Stevens, D. S. Lind, and B. Lok. 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, 2007.

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

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

[16] J. Nielsen and T. K. Landauer. 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, 1993.

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

[18] T. B. Pence, L. C. Dukes, L. F. Hodges, N. K. Meehan, and A. Johnson. 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, 2013.

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

[20] A. Rajak and K. Saxena. 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, 2010.

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

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

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

[24] D. Thalmann. 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, 2001.