Applying Mixed Reality to Simulate Vulnerable Populations for Practicing Clinical Communication Skills

Joon Hao Chuah, Benjamin Lok, *Member, IEEE*, and Erik Black

(a) User's perspective (b) Third-person perspective

Fig. 1. Mixed reality pediatric developmental exam allows a medical student to practice interviewing a parent and assessing a child. Note, the displays align correctly from the user's perspective (left) but incorrectly from a third-party perspective (right).

Abstract—Health sciences students often practice and are evaluated on interview and exam skills by working with standardized patients (people that role play having a disease or condition). However, standardized patients do not exist for certain vulnerable populations such as children and the intellectually disabled. As a result, students receive little to no exposure to vulnerable populations before becoming working professionals. To address this problem and thereby increase exposure to vulnerable populations, we propose using virtual humans to simulate members of vulnerable populations. We created a mixed reality pediatric patient that allowed students to practice pediatric developmental exams. Practicing several exams is necessary for students to understand how to properly interact with and correctly assess a variety of children. Practice also increases a student's confidence in performing the exam. Effective practice requires students to treat the virtual child realistically. Treating the child realistically might be affected by how the student and virtual child physically interact, so we created two object interaction interfaces - a natural interface and a mouse-based interface. We tested the complete mixed reality exam and also compared the two object interaction interfaces in a within-subjects user study with 22 participants. Our results showed that the participants accepted the virtual child as a child and treated it realistically. Participants also preferred the natural interface, but the interface did not affect how realistically participants treated the virtual child.

Index Terms—Virtual humans, medical education, social presence, presence

1 Introduction

Pediatric developmental exams are important for early intervention which can substantially improve a child's life outcomes [14]. However, exposing health science students to a wide variety of pediatric cases is difficult due to the impracticality and unethical nature of subjecting a child to several exams by different students. To solve this problem, we propose using virtual children to simulate these exams, and we created a mixed reality (MR) pediatric patient that allowed students to practice developmental exams.

In health sciences education, students often practice interview and exam skills by working with standardized patients (people that role play having a disease or condition) [1]. However, standardized patients either do not exist or have extremely limited availability for vulnerable populations such as children, the physical disabled, and the intellectually disabled [23, 26]. These vulnerable populations can be challenging to work with in the real world and often require special attention and care. Yet, students receive few opportunities to practice working with them before becoming working professionals [26].

Experience examining a wide range of pediatric patients is necessary because pediatric developmental exams are an ill-structured domain [21]. An ill-structured domain is one where the cases are nominally of the same type but vary substantially, and as a result mastering the domain requires extensive exposure to a variety of cases. A variety of children can be simulated, providing students with the practice they need in a safe, standardized environment.

The MR pediatric developmental exam builds upon prior virtual-human-based training systems [11, 16]. These prior systems use conversational virtual humans to train users on interpersonal skills such as interviewing a patient. To leverage this prior work, we chose to focus on pediatric exams where the child is old enough to talk. Four-year-

- *Joon Hao Chuah is with University of Florida. E-mail: jchuah@cise.ufl.edu.*
- *Benjamin Lok is with University of Florida. E-mail: lok@cise.ufl.edu.*
- *Erik Black is with University of Florida. E-mail: ewblack@ufl.edu.*

Manuscript received 13 September 2012; accepted 10 January 2013; posted online 16 March 2013; mailed on 16 May 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

1077-2626/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

old children are old enough to communicate in simple but complete sentences. Exams on four-year-old children include asking the child questions and also asking the child to copy shapes on a sheet of paper.

For the MR pediatric developmental exam to be successful, we believed users needed to perceive and treat the virtual humans realistically. To create these virtual humans, we leveraged our prior work on creating socially present virtual humans [6]. These virtual humans use MR technology to share the physical space with the user. Sharing the physical space increases the user's sense that the virtual humans are real people and elicits realistic behavior towards the virtual humans. We made the hypotheses that MR children could elicit realistic user behavior and that the MR exam could provide an educational benefit.

We also made a secondary hypothesis that users would treat the virtual humans more realistically if both the virtual humans and the user could directly interact with shared objects in the MR space. To test this secondary hypothesis, we created two interfaces that replicated the functions of drawing on paper. One interface used an MR sheet of paper that both the child and examiner drew on directly (Figure 1). The other interface used a mouse-based application where the examiner clicked buttons to draw shapes and see the child's drawings.

We evaluated our MR pediatric developmental exam in a within-subjects user study with 22 participants. Our primary goal was to determine if the MR exam was an effective and usable training tool. We also investigated if the object interaction interface affected how realistically participants perceived and treated the virtual humans. Our results showed the MR exam was an effective tool for inexperienced learners. All users treated the virtual humans realistically and the object interaction interface had no effect on perception of the MR child.

2 Related Work

2.1 Mixed Reality Virtual Humans

Prior research has shown MR technology can increase feelings of presence and elicit more realistic behavior towards virtual humans. Dow et al. created an augmented reality version of Facade [9]. They used a see-through head-mounted display to render life-size human virtual human in a physical apartment. Compared to virtual reality versions, they found the augmented reality version had more presence.

Johnsen et al. compared a life-size virtual human to a virtual human on a 22" monitor [12]. They found users interacted more engaged, empathetically, and naturally with the life-size virtual human.

Our own prior work focused on MR virtual humans that share the physical space with the user [6]. Compared with virtual humans projected on a wall, the MR virtual humans received higher self-report social presence ratings. Social presence is the feeling that another person is present in the room with you. We also found virtual humans projected on the wall elicited unrealistic behavior from the users but the MR virtual humans did not.

2.2 Virtual Patients

Virtual patients are virtual humans that learners can practice exam and interview skills on [13]. Virtual patients serve a similar function to standardized patients. Virtual patients offer opportunities for increased exposure both in terms of numbers and types of cases seen.

Virtual humans can simulate pathology that standardized patients cannot. Kotranza et al. simulated a patient with cranial nerve damage [15]. Cranial nerve damage can manifest as partial-paralysis of the face or the inability to move one eye past a certain point. Both of these symptoms can be accurately simulated with virtual humans but not with real humans. Similarly, our system simulates children, a population without standardized patients. Children cannot ethically be used for many hours of student training, nor can they reliably simulate problems such as developmental delays.

3 Background - Real Pediatric Developmental Exams

Multiple studies have determined that routine developmental screenings of children with a valid and reliable assessment lead to improved outcomes in children [24, 5]. Despite this data and the general recognition that developmental assessment is an important component of primary care pediatrics, health science practitioners have frequently reported low levels of self-efficacy related to the task [22].

Though many different developmental screeners are currently in use in different clinical situations, three common screeners include: the Denver II [8], the Ages and Stages Questionnaire (ASQ) [2] and the Parents' Evaluation of Developmental Status (PEDS) [20]. Each screener has different psychometric properties, but all three have been recognized as valid and reliable for a general pediatric population [22].

3.1 Current Training Methods

Few methods have been empirically established as an effective means of teaching health care professionals how to administer an assessment and discuss the results with family members [19]. According to educators and students we interviewed, current training methods focus mainly on teaching students what the developmental milestones are (e.g., four-year-old children should be able to draw circles). Most students receive little to no training on how to perform the exam prior to their clinical training rotations (hands-on training in real clinics).

Some students practice developmental exams on the children of friends or family. This provides valuable experience but it often lacks the challenges associated with exams on real children in clinic. Most importantly, the children of friends and family typically do not hail from economically-disadvantaged backgrounds. More often than not, these children are free of developmental delays.

3.2 Skills Tests for Four-Year-Old-Children

The MR pediatric developmental exam focused on skills for four-year-old children. Four-year-old children are of particular interest because their exams focus heavily on personal-social, language, and fine-motor skills. Testing these skills requires both verbal and non-verbal communication, and the results are harder to interpret than exams on infants. Examiners must learn not only how to properly administer the tests but also how to properly score the results.

Personal-social skills include items such as preparing cereal and getting dressed with no help. Language skills include speaking in understandable, complete sentences and naming colors. Assessing both these skills categories requires the examiner to ask the child questions. If the child is shy and answers "I don't know," the examiner may engage the child more or ask the parent if the child can perform the skill.

Assessing fine-motor skills involves extensive interactions with a pen and paper. The examiner typically draws a simple shape, such as a circle, and asks the child to name the shape. The examiner then passes the pen and paper to the child and asks the child to copy the shape. This process repeats with other shapes such as squares and crosses. Another test requires the child to identify the longer of two lines.

The fine-motor skills exam also requires the child to draw a person with four to six parts (e.g. head, eyes, arms, and legs). The examiner will point to parts and ask the child to name or count them.

4 Mixed Reality Pediatric Developmental Exam

The MR exam's virtual child and parent leveraged our prior work creating conversational, MR, virtual humans [6]. We expanded upon this work by creating two different object interaction interfaces. The object interaction interfaces allowed the examiner and child to draw shapes on paper. One interface allowed both the examiner and child to naturally, directly manipulate an MR sheet of paper. The other interface allowed the examiner and child to abstractly draw shapes using a mouse-based application.

4.1 Mixed Reality Virtual Humans

The MR exam sought for examiners to treat the virtual humans realistically. To achieve this goal, the exam used detailed meshes, real audio, MR technology, and Wizard of Oz control.

4.1.1 Graphics and Audio

Health science practitioners constantly look and listen for subtle problems with patients. Practitioners may diagnose a medical problem if the virtual human has low-fidelity animations and speech. An early version of the virtual child had low-fidelity, jerky animation, and during pilot testing pediatricians thought the child had muscular issues.

The final versions of the virtual humans interacted using a set of high-fidelity, pre-programmed animations and speeches. To ensure high-fidelity animations (e.g. drawing shapes on paper), we hired a professional animator. For high quality, natural speech, we pre-recorded real humans reading all of the virtual humans' pre-programmed speeches. Different humans recorded each virtual human. Each virtual child included 53 speeches, and each virtual parent included 42 speeches. The virtual child recordings were performed by the four and five-year old children of one author and a colleague. Recording each child took less than twenty minutes.

The virtual humans were all 60k triangle meshes generated using the Evolver.com website. Because Evolver did not have child meshes, a thin, short (140cm), young-looking adult mesh was used as a base for the child. The base mesh's bones were shortened so the resulting mesh had an appropriate height, torso length, arm length, and leg length.

4.1.2 Sharing the Physical Space

To encourage realistic behavior from the user, the virtual humans created the impression of sharing the physical space with the user. The child and parent were displayed as standing at life-size on a 24" LCD monitor and 40" LCD TV respectively. We based the standing setup on a recording of a real exam where all parties stood. However, users and medical experts later agreed sitting is more common, and we will correct this in the future.

The displays use head tracking and fish-tank VR [25] to provide 3D perspective. A Microsoft Kinect tracks the user's head position at approximately 30 hertz. As the user moves, the virtual camera also moves. This allows the user to see different parts of the virtual humans and makes the virtual humans feel more three-dimensional.

Combining fish-tank VR with a model of the real room creates the illusion of see-through displays. Fish-tank VR alters the virtual camera's perspective matrix to match the user's current position. This alteration warps the image so that the model of the room always remains visually aligned with the real room. The model of the real room was created in Google SketchUp and used photos as textures. The model and display positions were manually calibrated using a tape measure.

The virtual humans showed awareness of the user's physical position by maintaining eye contact. The virtual humans constantly turned their heads and eyes to follow the user as he moved about the room. The virtual child only broke eye contact to draw shapes on the paper.

4.1.3 Wizard of Oz

We used a Wizard of Oz control interface to avoid problems with artificial intelligence and also allow a human operator to take on a variety of roles. Artificial intelligence systems sometimes respond incorrectly because participants may have accents or phrase questions in unexpected ways. Incorrect responses could lead to participants having varying quality experiences. To avoid this problem, a human operator, rather than a computer, remotely controlled the characters. This technique, known as Wizard of Oz, is commonly used in studies focused on interfaces and aspects of virtual human presentation [7, 9, 18]. However, the participants were told the characters were computer-controlled because research shows people rate avatars higher for social presence if they believe the avatar is controlled by a human [10].

The human operator listened directly to the participant. The human operator could not physically leave the room for logistic reasons. However, the operator sat outside of the participant's field of view and told the participant he was there only to troubleshoot. When the participant finished speaking, the operator picked an appropriate response from a list of pre-programmed responses. The responses were displayed on a tablet (a Motorola Xoom), allowing the operator to silently select a response. The same human operator controlled the virtual humans for all participants. This human operator had extensive knowledge of the pre-programmed responses. The human operator, with the aid of medical practitioners, created all of the responses. The human operator also practiced with over ten pilot test participants.

If the participant asked the father a question without a pre-programmed response, the father answered "My wife would know the answer to that" or "My wife usually does that." If the participant asked the child a question without a pre-programmed response, the child responded "I don't know" or shrugged its shoulders. This is natural shy behavior for a child, and students must learn to rephrase questions to draw out the answer (e.g. "Did mommy help you get dressed this morning?") or redirect the question to the parent. If the participant asked the child to perform an action it was not pre-programmed to do (e.g. draw a letter), the child responded "No." Refusing to do something is also natural behavior for a child, especially when interacting with unfamiliar people.

4.2 Object Interaction Interfaces

The pediatric developmental exam for four-year-old children involves frequent interactions with a sheet of paper. The exam uses the paper to test the child's abilities to:

- Name and copy shapes drawn by the examiner
- Point to the longer of two lines drawn by the examiner
- Draw a person and name and count the body parts

We created two object interaction interfaces to test how perception of the virtual child was affected by the virtual child's ability or inability to interact with objects. One interface was a natural interface where both the virtual child and human user directly interacted with an MR sheet of paper (Figure 2). The other interface was a mouse-based interface where the virtual child and human user abstractly interacted with the paper through button clicks (Figure 3). Both interfaces used the same MR virtual humans and enabled both the examiner and child to communicate which shapes they drew.

4.2.1 Natural Interface

The natural interface allowed both the virtual child and the examiner (a real human user) to directly interact with a shared sheet of MR paper. The MR paper spanned two displays (Figure 2). The first display, a 10.1" Asus Transformer tablet, was placed flat on the table and displayed one half of the paper. The paper visually continued onto a 17" LCD monitor placed vertically on the edge of the tablet. The monitor's vertical orientation, coupled with fish-tank VR, allowed the monitor to correctly render the child's hand when he drew on the paper. A purely horizontal display would have resulted in a visual disconnect between the child's arm on the vertical 24" LCD and hand on the paper.

The user physically interacted with the paper using the tablet and a capacitive stylus. To draw shapes, the user traced dashed outlines of the shapes with the stylus. The outlines consisted of two squares, two circles, two crosses, a long line, and a short line. Tracing outlines was used instead of free-form drawing to keep the interface equivalent to the mouse-based interface. The mouse-based interface displayed buttons for drawing each shape, which gave the users a reminder of which shapes to draw. The buttons also prevented the user from drawing shapes the system would not recognize (e.g. triangles or letters).

Both the child and the user slid the paper back and forth. The child slid the paper towards himself to point to the longer of two lines drawn by the user. The user slid the paper closer to herself to point at body parts in the child's drawing. The user slid the paper closer to herself to get a better look at the child's drawings or slid the paper towards the child to let him draw on it. To slide the paper, the user slid two fingers on the tablet. This motion mimiced using two fingers to slide a real piece of paper in a straight line.

When the child's drawing of a person slid off the monitor and on to the tablet, shaded circles appeared over body parts (similar to Figure 3). These circles indicated hot zones where the user could point using the stylus. The user could point to a body part and ask the child to name the body part or color in the drawing. The user could also ask the child to count the body parts as the user tapped them one by one.

4.2.2 Mouse-Based Interface

The mouse-based interface abstracted the object interaction into buttons in a mouse-based application (Figure 3). The mouse-based application communicated with the rest of the system but was not integrated

Fig. 2. Virtual child and MR paper. The child spans the top and middle displays (both LCD monitors), and the paper spans the middle and bottom displays (an LCD monitor and a tablet). Fish-tank VR ensures the child and paper remain visually consistent across displays.

Fig. 3. Paper application. The left half contains buttons for the user to draw shapes. The right half displays the child's drawings. The button in the bottom right appears when the child draws a shape. Clicking the bottom right button will reveal the child's drawing. Note: Lower screen image simulated for visibility.

into the MR environment. The application was a C# application displayed on the same 17" LCD used for the MR paper, and the user controlled the application using a standard wireless mouse.

The application presented buttons for drawing each shape labeled 'Draw *shape*'. When the user clicked a button to draw a shape, the button was replaced by an image of the shape (Figure 3). The images were perfect representations of the shape rather than hand-drawn images. The perfect representations emphasized that the user was not directly manipulating an object.

When the child drew a shape, a button labeled 'Show Child's Drawing' appeared. Clicking the button replaced the button with an image of the child's drawing. This additional step emphasized the disconnect between the mouse-based interface and the MR environment. The child could not directly 'draw' on the paper application. However, the child still made drawing motions, and the button did not appear until the child finished the drawing motion. Because the mouse-based interface did not involve direct manipulation of the paper, there was no equivalent for sliding the paper back and forth.

The position of the 17" LCD intentionally blocked the lower portion of the child. This prevented the user from seeing the child's hand when the child drew shapes. Seeing the child's hand drawing directly on the virtual table without virtual paper would be odd, and including the virtual portion of the paper would create the impression that the child could directly manipulate objects like in the natural interface.

5 User Evaluation

The user study's primary goals were to evaluate the MR pediatric developmental exam's effectiveness as a training tool, usability, and ability to elicit realistic treatment of the virtual humans. The study also examined the effect of the object interaction interface on perception of the virtual child as a real child. The study used a within-subjects design where participants performed two exams, one with each object interaction interface. To avoid participants assessing the same child more than once, we created two child-parent pairs. Both of these children had similar developmental statuses and both parents had similar concerns; only appearance, voice, and minor details differentiated the pairs. We counter-balanced both the order of object interaction interfaces and the order of child-parent pairs, resulting in four groups. We also created a third child-parent pair that instructed participants on how to use the system as a part of an interactive tutorial.

5.1 Participants

The user population for the evaluation consisted of health professions students with pediatric developmental exam training and pediatric residents. Participants were recruited by e-mail announcement over a period of three weeks. The announcement offered participants the opportunity to practice performing a developmental exam on a virtual child as part of a training exercise. Participation was voluntary, and participants were compensated with a $20 gift card. Some students also received extra credit for participating, but no participants were graded as part of the user evaluation.

In total 22 people participated in the evaluation (mean age 25.45, SD=5.11, 18 female). Ten participants were undergraduate nursing students, seven were medical students, three were pediatrics residents, one was a graduate nursing student, and one was a physician's assistant student. All but one had prior experience either performing or observing pediatric developmental exams.

Participants were evenly distributed into the four counter-balanced groups. All groups had similar gender and student-type compositions.

5.2 Procedure

All four groups followed identical study procedures:

1. Completed an informed consent form and were briefed about the goals and scope of the training exercise.
2. Interacted with the tutorial child-parent pair and learned to use one object interaction interface
3. Completed a demographics and background survey
4. Performed a developmental exam with one child-parent pair using one object interaction interface
5. Completed a feedback survey
6. Interacted with the tutorial child-parent pair and learned to use the other object interaction interface
7. Performed a developmental exam with the other child-parent pair using the other object interaction interface
8. Completed a feedback survey and a free-form feedback interview

5.2.1 Tutorials

The tutorials taught the participant how to interact with the virtual humans. The tutorials also gave the participant a chance to practice the object interaction interfaces. During the tutorial, a virtual parent instructed the participant how to verbally communicate and use one object interaction interface. The parent spoke the majority of the time and prompted the participant with what to say and do. The prompts were the same in all conditions except when referring specifically to the object interaction interface. The first tutorial had four phases:

1. Introduction - The parent first introduces himself and the child, then explains the goals and scope of the exercise.
2. Naming colors - The parent explains how to use the first object interaction interface to point at different colored parts in the child's drawing and ask the child to name the color.
3. Drawing shapes - The parent told the participant to draw a square, ask the child to name the shape, and ask the child to copy the shape. The child then copied the shape incorrectly (drawing a triangle instead of a square), and the parent told the participant to draw the shape again while telling the child what to pay attention to. Asking the child to repeat a failed test commonly occurs in developmental exams; only repeated failures should be considered true failures.
4. Closing - The parent instructed the participant how to ask both the parent and child questions, and how to deliver the exam results to the parent.

The second tutorial was a subset of the first tutorial and only included the naming colors and drawing shapes phases. Both tutorials used the same child-parent pair, and the tutorial child-parent pair was not used in either of the developmental exams.

5.2.2 Developmental Exams

The developmental exams followed four steps outlined explicitly on a sheet of paper given to the participants:

- Find out what the parent's concerns are
- Assess the child's personal-social, fine motor-adaptive, and language skills
- Deliver your assessment to the parent
- Address any questions and concerns the parent has

The parent began the exam by introducing himself and the child and expressing concern that the child is behind compared to other children. If the participant asked for more information, the parent would elaborate on his concerns with specific examples of things the child cannot do or is not good at. After interviewing the parent, participants went on to assessing the child's skills.

To help assess the child's skills, participants received a printed list of skills taken from two standard developmental screeners, the Denver II [8] and ASQ [2]. These skills include both fine-motor and language skills. For language skills, the sheet listed items to ask the child (e.g. phone number) but left the participant free to phrase the question however she desired. To pass the language skills test, the child should know at least four of the six items on this list. Both children only knew four of the items, and the children answered 'No' or 'I don't know' for the other two items. This creates an opportunity for ambiguity. Participants could interpret this as the child truly not knowing the answer, the child being shy, or the application making a mistake.

When the participant was done assessing the child, she delivered the results of her assessment to the parent. Both children were created to be in the 50th percentile and hence normal for development. However, some participants assessed one or both of the children as slightly behind. This happened when a participant incorrectly failed the child for a skill (e.g. failing a circle because it is not round enough when any enclosed form should pass). This also happened if the participant asked a large number unexpected questions and receive too many "I don't know" responses.

After receiving the results, each parent again expressed concern that the child was behind his peers (e.g., "My sister's kids all knew the alphabet at his age. Why doesn't he?"). Similar questions commonly arise in real exams, and examiners must learn to respond in a way that satisfies the parent. However, students do not typically receive instruction on this skill and may not have encountered this situation.

5.3 Metrics

The study evaluated the MR exam on three aspects: educational benefit, usability, and perception of the virtual humans. For all aspects, the study used both quantitative measures in the form of self-report surveys and qualitative measures in the form of post-exercise free-response questions.

5.3.1 Educational Benefit

Changes in a participant's confidence levels were used as indicators of the exam's educational benefit. Participants used survey items to rate their confidence on three exam skills: interviewing a parent, interacting with a child, and assessing a child's development. The survey items used a five-point scale ranging from 'Not at all confident' to 'Very confident.' The three survey items appeared in the pre-exercise, mid-exercise, and post-exercise surveys. The three survey items were

- (Interviewing) Interviewing a parent about his/her child's development.
- (Interacting) Interacting with the child during a developmental exam
- (Assessing) Assessing a child's development

Because experience level may play a role in confidence, the participants were divided into two groups for analysis - *inexperienced* and *experienced*. Groups were based on the estimated number of exams performed prior to the study. The mean estimate of eight exams was used as the dividing line between the groups. The inexperienced group contained 14 participants, all of whom estimated having performed five or less exams. The experienced group contained eight participants, all of whom estimated having performed 10 or more exams. The experienced group contained all three residents and five of the seven medical students.

Correctly assessing the child's developmental status was not used as a metric for two reasons. First, participants that performed a more thorough exam received more "I don't know" or "No" responses because they asked questions or gave instructions without preprogrammed responses. These responses influence the assessment and can be interpreted as a sign of developmental delays. Second, participants did not have access to a complete scoring sheet. Some participants were familiar with the Denver II scoring sheet, whereas others were familiar with the ASQ scoring sheet. For fairness, neither score sheet was provided, so participants had to form an assessment based solely on their own judgement.

5.3.2 Usability

Participants rated the system's usability using eight Likert-type survey items from the System Usability Scale [17]:

- Use Frequency - I think that I would like to use this system frequently
- Complexity - I found the system unnecessarily complex
- Ease - I thought the system was easy to use
- Integration - I found the various functions in this system were well integrated
- Inconsistency - I thought there was too much inconsistency in this system
- Learnability - I would imagine that most people would learn to use this system very quickly
- Awkwardness - I found the system awkward to use
- Confidence - I felt confident using the system

Participants also answered six survey items rating which interface was better. The items used a seven-point scale from 'Mouse was much better' to 'Tablet much better'. The six items were:

- Overall
- Natural conversation flow
- Practicing a developmental exam
- Assessing the shapes the child drew
- Communicating what to pay attention to (e.g. asking the child to try drawing a square again)
- Communicating which shape I wanted the child to draw

5.3.3 Perception of Virtual Humans

Both self-report survey items and observed behaviors measured how realistically participants perceived the child. For self-report survey items, the study used five Likert-type items adapted from Bailenson et al. [4]. Because the MR exam involved two virtual humans, the surveys asked each item twice and replaced the word 'person' with 'child' or 'parent.' The survey items were on a seven-point scale from 'Strongly disagree' to 'Strongly agree.' The five items, with abbreviations for reporting results in parentheses, were:

- (Presence) I perceive that I am in the presence of a child in the room with me.
- (Watching) I feel that the child is watching me and is aware of my presence.
- (NotReal) The thought that the child was not a real person crossed my mind often.

- (Sentient) The child appeared to be sentient, conscious, and alive to me.
- (Computerized) I perceived the child as being only a computerized image, not as a real person.

According to further research by Bailenson et al. [3], behavioral measures of social presence may be more sensitive than self-report survey items. The behavioral measure was the tone of voice the participant used. People typically talk to young children using a distinct, child-specific tone of voice. An audio-only recording of each participant speaking to the virtual child was played to five independent, third-party coders. Each coder answered the question "Is the adult using a tone of voice reserved for children (e.g., higher pitch, extra emphasis on certain words)?" with yes, no, or unsure. If at least three coders answered yes, then the participant was considered to have used the child-specific tone of voice.

Coders also listened for another sign of talking to a child, acknowledging what the child said. Each coder answered the question "Did the adult acknowledge the child's answers at least once (e.g., repeating the answer, saying good job)?" If less than four coders answered yes, then a first-party coder reviewed the recording as confirmation.

6 Results and Discussion

We report first on the results of the MR exam without considering child-parent pair or object interaction interface as factors because child-parent pair and object interaction interface were not correlated with statistically significant differences in any of the metrics. At the end of this section, we discuss qualitative diferences between the object interaction interfaces.

6.1 Educational Benefit

6.1.1 Educational Benefit Results

The confidence scores showed the *inexperienced* and *experienced* groups started with differences, but the difference shrank after using the MR exam (Figure 4). In the inexperienced group, 71% of participants reported a confidence change for at least one of the three survey items, and confidence changes were common for all three survey items. In the experienced group, confidence changes were common for only one survey item.

Pre-exercise, the experienced group had statistically higher scores for Interviewing ($p < 0.05$) and Interacting ($p < 0.05$), but the difference was not statistically significant for Assessing ($p = 0.05$) according to a one-tailed T-test. Post-exercise, the differences were no longer statistically significant for any of the items (Interviewing $p = 0.09$, Interacting $p = 0.14$, Assessing $p = 0.44$).

In the inexperienced group, some participants confidence increased while others decreased. Eight of the 14 inexperienced participants reported a confidence increase for at least one item and four of the 14 reported a confidence decrease for at least one item (two reported an increase for one item and a decrease for another item). Only four reported no confidence change for all items. Confidence changes were about equally common in all three survey items (five in Interviewing, seven each in Interacting and Assessing). However, as a whole, the inexperienced group did not show a significant change for any of the three items according to a repeated-measures ANOVA analysis.

In the experienced group, 63% of the confidence changes were for Assessing. Three participants' self-reported scores for Assessing decreased and two increased. The Interviewing item had only one participant change (a decrease), and the Interacting item had only two participants change (both increases).

6.1.2 Educational Benefit Discussion

The results suggest the MR exam provided an educational benefit to most of the *inexperienced* group and some of the *experienced* group. In the inexperienced group, 71% showed some change in confidence. The change was not statistically significant because some participants' confidence increased while others decreased. This variance could be because some students were well-prepared but under confident while other students were less-prepared but over confident. For well-prepared students, the additional practice opportunity increased their confidence. For less-prepared students, the MR exam served as a wake up call emphasizing the areas that needed practice. This interpretation is supported by free-response feedback. One participant said the MR exam was a 'Good idea because I didn't realize how little I knew about peds assessment." Another said "I think it would be very helpful because I'm scared to death of working with pediatric patients... it would be great to be able to sit in here for an hour."

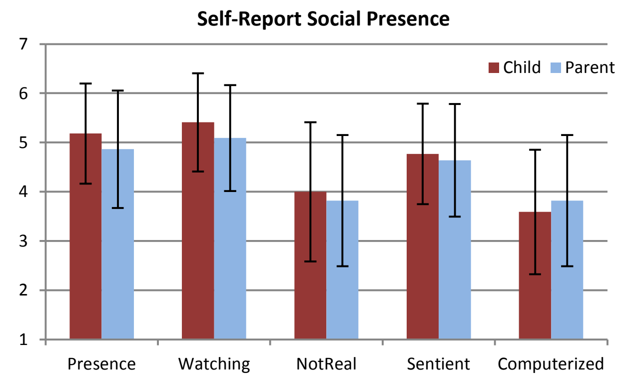

Fig. 5. Self-report social presence ratings (mean and standard deviation) for the child (red, left) and parent (blue, right). For the negative statements NotReal and Computerized, lower is better.

For members of the experienced group, the MR exam could serve as a refresher on developmental exams. Changes in confidence were mainly seen for 'Assessing a child's development' which suggests some participants realized they had forgotten details of developmental exams. Confidence in communication skills (Interviewing and Interacting) started high and remained high because these are measures of interpersonal skills, not exam skills.

6.2 Usability

6.2.1 Usability Results

The MR exam received mean usability scores in the range of 'Slightly agree' to 'Agree' for all positive statements and 'Slightly disagree' to 'Disagree' for all negative statements (Use frequently 5.45 ± 1.07, Complexity 2.20 ± 1.00, Ease 5.98 ± 0.79, Integration 6.00 ± 0.68, Inconsistency 2.39 ± 1.24, Learnability 6.05 ± 1.03, Awkwardness 3.07 ± 1.59, Confidence 5.50 ± 1.06).

6.2.2 Usability Discussion

The MR exam's high usability likely contributed to participants feeling the MR exam was educationally beneficial and something they would like to use again. Usability is important when presenting new, unfamiliar technology to people. Many people have difficulty with computers, and these people can be intimidated by relatively cutting-edge technology such as mixed reality and virtual humans. This intimidation barrier was overcome by the MR exam, as shown by the high usability scores. One participant noted she was "Not computer savvy at all, so if I can figure it out, anyone can figure it out."

6.3 Perception of Virtual Humans

6.3.1 Perception of Virtual Humans Results

Overall, both the child and parent received average ratings of neutral or better for all self-report social presence survey items (Figure 5). While the scores for the child and parent were similar, free-response feedback showed some participants perceived the child and parent differently.

In 71% of the exam recordings, the participant talked to the virtual child using the child-specific tone of voice, and in 95% of the recordings, the participant acknowledged the child's answers. Only four of the 22 participants never used the child-specific tone of voice.

Some participants took extra steps to engage and encourage the child. For example, one participant asked the child "I like your outfit you have on. Were you able to put that on yourself?" instead of simply asking the parent "Can he dress himself in the morning?"

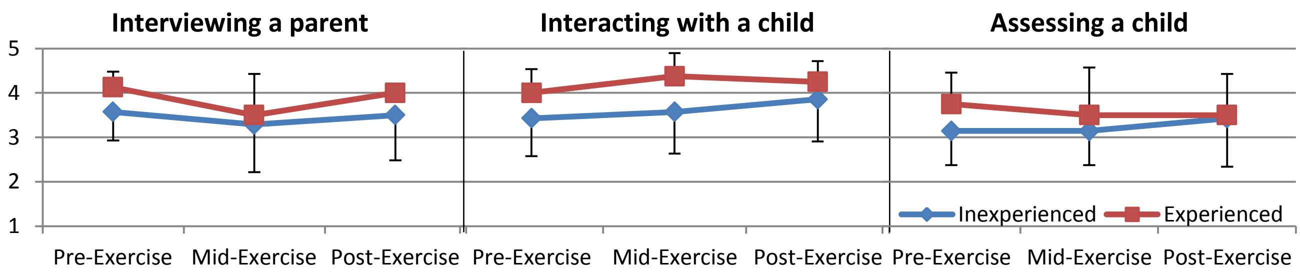

Fig. 4. Average confidence levels for the experienced and inexperienced groups. For readability, standard deviations are shown as positive error bars for the experienced group and negative error bars for the inexperienced group.

When the child refused to perform a test or answered "no" to a question, participants interpreted this in two different ways. Some attributed the refusal to the child's personality, saying "Since he's not totally cooperative, it's realistic." Other participants blamed the computer, making comments such as "I didn't know if he didn't answer because I asked wrong or he was programmed to answer it that way."

Several participants with developmental exam experience considered the child more realistic than the parent. Participants commented not on the parent's appearance but instead on his behavior. Specifically, the parent never interjected in the exam and overall seemed less involved than a real parent. One experienced participant commented "The parent needs a little work, not as realistic. The kid was good, very interactive... Usually parents are more overbearing, something you have to work with in real life." Another expected "Parents would be able to give longer answers or follow up questions. Also, a lot of parents will encourage the child during the examination process."

6.3.2 Perception of Virtual Humans Discussion

The MR exam successfully created virtual humans that users perceived and treated realistically. Most importantly, the MR exam successfully simulated children, a vulnerable population. The virtual children elicited realistic behavior from the participants. Participants constantly acknowledged and encouraged the virtual child like they would a real child. Participants received no instructions to talk to the child this way, and they could have treated the interaction as a pure information gathering exercise. Instead of just asking a series of questions and giving instructions, participants tried to engage the child. This engagement behavior was observed even in the four participants that did not talk to either virtual child in the child-specific tone of voice.

The varying responses to the child responding "No" to questions and instructions emphasizes an important point when simulating humans, especially children. Users may be uncertain whether they should attribute refusals to intended behavior or to a computer error. This ambiguity could be alleviated by providing the user with feedback that the system understood the question. For example, when the child refuses to draw a shape the parent could add "I think he's tired of drawing. Maybe you should try something else."

The feedback comparing the realism of the virtual child and virtual parent shows the importance of behavioral realism. Even though the self-report social presence ratings for both virtual humans were similar, participants felt there was a difference. These findings on behavioral realism reflect similar findings in our previous work where one virtual human's unrealistic behavior made it seem less realistic than another virtual human [6].

6.4 Object Interaction Interfaces

6.4.1 Object Interaction Interfaces Results

Most participants expressed a preference for the natural interface (Figure 6), but the object interaction effect did not show statistically significant effects on perception and treatment of the virtual humans.

For the child-specific tone of voice, four of the 22 participants only used the child-specific tone of voice with the natural interface. Twelve of the 22 participants used the child-specific tone of voice with both

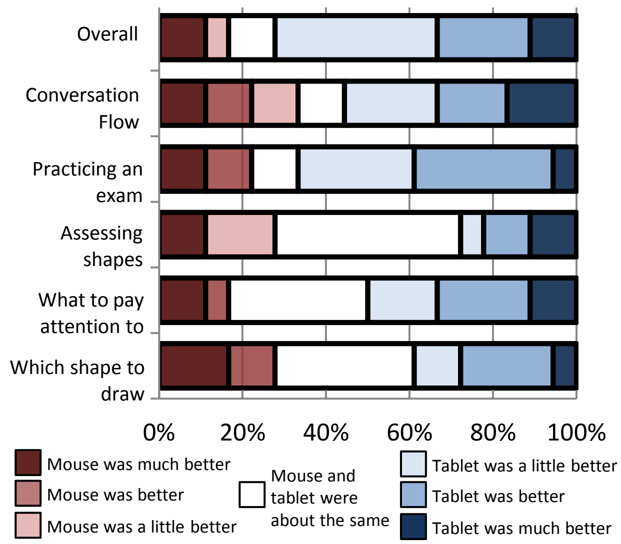

Fig. 6. Participant preference for the mouse-based interface (left, red shades) or natural interface (right, blue shades). Each bar shows the percent of participants that chose that response. Note, only 18 of the 22 participants answered these survey questions.

interfaces, and four used the child-specific tone of voice with neither interface. The remaining two used the child-specific tone of voice with the mouse-based interface, but the recordings of their natural interface interactions were lost due to technical problems.

Self-report social presence scores showed no statistically significant differences between the natural and mouse-based interfaces for the child according to a repeated-measures ANOVA (Table 1).

Table 1. Self-report social presence scores (mean and standard deviation) for the virtual child in the natural and mouse-based conditions. * indicates negative items.

	Natural	Mouse-Based	p
Presence	5.00 ± 1.07	5.36 ± 0.95	0.07
Watching	5.50 ± 0.91	5.32 ± 1.09	0.30
NotReal*	3.91 ± 1.48	4.09 ± 1.38	0.51
Sentient	4.62 ± 1.12	4.86 ± 0.91	0.10
Computerized*	3.64 ± 1.36	3.55 ± 1.18	0.69

6.4.2 Object Interaction Interfaces Discussion

Overall, it appears that the natural interface caused participants to report a more realistic exam experience, but the natural interface did not make participants treat the virtual child more realistically. Extra effort should be spent not on making object interactions more realistic and physical but instead on making virtual human behavior more realistic.

The results suggest the natural interface is more realistic, but both object interaction interfaces are effective communication channels. Over half the participants rated the natural interface better for survey items related to realism - practicing an exam, conversation flow, and overall (Figure 6). The increased exam realism is supported by participant feedback such as "Drawing it makes you feel more comfortable as a pediatrician, feel like I'm actually administering the test as opposed to clicking a button." However, participants were about evenly divided for items related to communicating concrete details - assessing shapes, what to pay attention to, and which shapes to draw (Figure 6).

It appears increasing the object interaction realism mattered less than behavioral realism. Because the child behaved realistically, participants tended to perceive and treat the child realistically regardless of object interaction realism. It is also possible that making the virtual humans share the physical space with the user was sufficient to create a high degree of social presence and elicit realistic user behavior.

7 Other Vulnerable Population Simulations

Our results suggest virtual humans can successfully simulate one case with a vulnerable population, a healthy pediatrics patient. Within pediatrics there are many other potential cases, and there are also other vulnerable populations outside of pediatrics.

Within pediatrics, the MR developmental exam opens the possibility of exposing students to sensitive cases such as signs of child abuse. Potential child abuse cases are, unfortunately, encountered with real children, and examiners must practice properly handling the situation. Examiners must also learn to distinguish between signs of child abuse or neglect and normal childhood accidents. Such cases can be safely practiced with virtual children and virtual parents.

Outside of pediatrics, there are other vulnerable populations such as the intellectually disabled. An intellectually disabled person could possibly be simulated by a standardized patient with especially good acting skills, but most standardized patients are not trained actors. However, one good actor could be multiplied using MR virtual human technology. If a good actor is recorded once, her voice could be reused hundreds of times as part of a virtual human.

The Wizard-of-Oz control technique can leverage the existing large pool of trained standardized patients. Wizard-of-Oz control allows standardized patients to play roles that do not match their appearance. For example, a twenty-year-old, healthy, female standardized patient can control a fifty-year-old, male virtual patient with an amputated leg. The standardized patient can easily take on this disparate role because the virtual patient has a set of pre-programmed responses with high-fidelity audio and animation.

8 Conclusion

We successfully applied MR technology to simulate a medical exam with a vulnerable population. The medical exam, a pediatric developmental exam, involved complex verbal and non-verbal interactions with both a child and a parent. The child and parent both elicited realistic reactions from the users. Users tried to engage the child and encourage him to talk and do more. Users also tried to assuage the parent's concerns. Our results suggested an MR exam could allow learners to practice exams and experience exposure to a wider variety of children and parents than they currently encounter.

Acknowledgments

The authors wish to thank Drs. Meredith Mowitz, Lindsay Thompson, and Maria Kelly for their assistance developing accurate pediatric cases. This work was supported in part by a University of Florida Alumni Award, Children's Miracle Network, and NSF IIS-0803652.

References

[1] M. A. Ainsworth, L. P. Rogers, J. F. Markus, N. K. Dorsey, T. A. Blackwell, and E. R. Petrusa. Standardized Patient Encounters: A Method for Teaching and Evaluation. *JAMA: The Journal of the American Medical Association*, 266(10):1390–1396, Sept. 1991.

[2] Ages & Stages Questionnaires. http://agesandstages.com/. Sept. 7, 2012.

[3] J. Bailenson, E. Aharoni, A. Beall, R. Guadagno, A. Dimov, and J. Blascovich. Comparing Behavioral and Self-Report Measures of Embodied Agents Social Presence in Immersive Virtual Environments. In *7th Annual International Workshop on Presence, Valencia*. Citeseer, 2004.

[4] J. Bailenson, J. Blascovich, A. Beall, and J. Loomis. Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, 29(7):819, 2003.

[5] D. B. Bailey Jr., K. Hebbeler, D. Spiker, A. Scarborough, S. Mallik, and L. Nelson. Thirty-six-month outcomes for families of children who have disabilities and participated in early intervention. *Pediatrics*, 116(6):1155, 2005.

[6] J. H. Chuah, A. Robb, C. White, A. Wendling, S. Lampotang, R. Kopper, and B. Lok. Increasing agent physicality to raise social presence and elicit realistic behavior. In *2012 IEEE VR*, pages 19–22. IEEE, Mar. 2012.

[7] N. Dahlback, A. Jonsson, and L. Ahrenberg. Wizard of Oz studies why and how. *Knowledge-Based Systems*, 6(4):258–266, Dec. 1993.

[8] Denver Developmental Materials, Inc. http://denverii.com/. Sept. 7, 2012.

[9] S. Dow, M. Mehta, E. Harmon, B. MacIntyre, and M. Mateas. Presence and engagement in an interactive drama. In *Proceedings of CHI 07*, pages 1475–1484. ACM Press, 2007.

[10] R. Guadagno, J. Blascovich, J. Bailenson, and C. Mccall. Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology*, 10(1):1–22, 2007.

[11] R. W. Hill, Jr, A. W. Hill, J. Gratch, S. Marsella, J. Rickel, W. Swartout, and D. Traum. Virtual Humans in the Mission Rehearsal Exercise System. *In KI Embodied Conversational Agents*, 17:32–38, 2003.

[12] K. Johnsen, D. Beck, and B. Lok. The Impact of a Mixed Reality Display Configuration on User Behavior with a Virtual Human. In *Intelligent Virtual Agents*, pages 42–48, 2010.

[13] K. Johnsen, A. Raij, A. Stevens, D. Lind, and B. Lok. The validity of a virtual human experience for interpersonal skills education. In *Proceedings of CHI 07*, pages 1049–1058. ACM, 2007.

[14] L. Karoly. *Investing in our children: What we know and don't know about the costs and benefits of early childhood interventions*. Rand, Jan. 1998.

[15] A. Kotranza, K. Johnsen, J. Cendan, B. Miller, D. Lind, and B. Lok. Virtual multi-tools for hand and tool-based interaction with life-size virtual human agents. In *2009 IEEE 3DUI*, pages 23–30. IEEE, 2009.

[16] A. Kotranza, B. Lok, C. Pugh, and D. Lind. Virtual Humans That Touch Back: Enhancing Nonverbal Communication with Virtual Humans through Bidirectional Touch. In *2009 IEEE VR*, pages 175–178. IEEE.

[17] J. Lewis. The factor structure of the system usability scale. *Human Centered Design*, pages 1–10, 2009.

[18] D. Maulsby, S. Greenberg, and R. Mander. Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of CHI '93*, pages 277–284, New York, New York, USA, May 1993. ACM Press.

[19] P. Nicol. Using the Ages and Stages Questionnaire to teach medical students developmental assessment: a descriptive analysis. *BMC Medical Education*, 6:29, 2006.

[20] Parents' Evaluation of Developmental Status. http://pedstest.com/. Sept. 7, 2012.

[21] R. Spiro, P. J. Feltovich, M. Jacobson, and R. L. Coulson. Cognitive Flexibility , Constructivism , and Hypertext : Random Access Instruction for Advanced Knowledge Acquisition in Ill-Structured Domains Rand J . Spiro Center for the Study of Reading University of Illinois School of Medicine Michael J . Jacobson C. *Constructivism and the technology of instruction: A conversation*, pages 57–75, 1992.

[22] L. a. Thompson, S. Y. Tuli, H. Saliba, M. DiPietro, and J. a. Nackashi. Improving developmental screening in pediatric resident education. *Clinical pediatrics*, 49(8):737–42, Aug. 2010.

[23] T.-C. Tsai. Using children as standardised patients for assessing clinical competence in paediatrics. *Archives of disease in childhood*, 89(12):1117–20, Dec. 2004.

[24] H. M. E. Van Agt, H. A. Van Der Stege, H. De Ridder-Sluiter, L. T. W. Verhoeven, and H. J. De Koning. A cluster-randomized trial of screening for language delay in toddlers: effects on school performance and language development at age 8. *Pediatrics*, 120(6):1317–1325, 2007.

[25] C. Ware, K. Arthur, and K. S. Booth. Fish tank virtual reality. In *Proceedings of CHI 93*, pages 37–42. ACM Press, 1993.

[26] L. J. Woodard, S. M. Havercamp, K. K. Zwygart, and E. a. Perkins. An innovative clerkship module focused on patients with disabilities. *Academic medicine*, 87(4):537–42, Apr. 2012.