Lessons Learned in Assessing Human-Virtual Human Interaction

Catherine Zanbaka, Paula Goolkasian, Larry F. Hodges

University of North Carolina at Charlotte
9201 University City Boulevard
Charlotte, NC 28223
{czanbaka, pagoolka, lfhodges}@uncc.edu

Benjamin Lok

University of Florida
P.O. Box 116125
Gainesville, FL 32611
lok@cise.ufl.edu

Abstract

To avoid the "file drawer effect" of unreported null results, we report on an experimental study in assessing human-virtual human interaction. In this study, participants were asked to spend about ten minutes with a projected life-sized virtual human. The virtual human interacted with the participants by speaking to them and exhibited three different tones of voice and facial expressions corresponding to a *happy*, *sleepy*, or *grumpy* personality. The virtual human asked participants to help her with a visual memory task where they responded, via keyboard, to a series of question sets about various pictures. We hypothesized that the virtual human's personality will have an effect on the amount of time participants are willing to help the virtual human on a task. In the results of our preliminary experiment, the personality of the virtual human had no effect on interaction time. First, we summarize the study and results, next we discuss some of the probable causes of our results, and finally we outline the lessons learned.

1 Introduction

1.1 Motivation

A common problem in many fields of research is what is known as the "file drawer effect" of unreported null results. The file drawer effect refers to the tendency to publish research with positive outcomes much more frequently than research with negative outcomes (Scargle, 2000). A consequence of this effect is that some researchers may duplicate efforts in testing a particular hypothesis.

The experimental study of human-virtual human interaction is a relatively new research area, making it more likely that some experiments will not be a success. Several challenges currently face virtual human researchers, including multimodal interaction, real-time animation, artificial intelligence, as well as assessing the effectiveness of the use of virtual humans. It is also unclear if interaction with virtual humans should be compared to traditional interfaces or to interaction with real humans. These challenges make it essential for researchers to publish their results so that others will learn from their experiences.

Virtual humans have a wide range of potential uses, including helping people with social and psychological conditions, medical training, diversity training, tele-collaboration, and enhancing virtual environments. However, a challenge for researchers is determining whether or not virtual humans can be used effectively in these areas. In order to successfully exploit virtual humans for applications where social interaction plays an important role, researchers must first determine if there exists a measurable similarity in responses to virtual and real humans. The purpose of our research is to experimentally determine the characteristics necessary in a computer generated and controlled virtual human such that the social and psychological responses of a real person to the virtual human are similar to the responses evoked by a real human. We explored the effects of virtual human personality on time spent interacting with the virtual human.

1.2 Related Work

Researchers have investigated how people respond to computers and virtual humans. For instance, Nass and Moon (2000) have shown that people react to and attribute very human characteristics to computers, such as the computer's helpfulness, expertise, and friendliness. Mel Slater's group at UCL has also conducted studies on the effects and social ramifications of having avatars in virtual environments. They were able to elicit emotions such as

embarrassment, irritation, and self-awareness in virtual meetings. They found that the presence of avatars was important for social interaction, task performance, and presence (Slater & Usoh, 1994).

Virtual humans have also been used in training. Slater, et al. (2000) examined the extent to which virtual humans could be used by actors and a director to rehearse for a live performance. The authors suggest that a performance level was reached in the virtual rehearsal which led to a successful live performance, one that could not have been achieved by only learning of lines or video conferencing. Virtual humans are also being used to allow medical students to experience the interaction between a patient and a medical doctor using natural methods of interaction with a high level of immersion (Johnsen et al., 2005).

Furthermore, researchers have applied social psychology to virtual human research. It has been shown that exposure to a virtual audience can be helpful in the treatment of the fear of public speaking (Anderson, Rothbaum, & Hodges, 2003). The Research Center for Virtual Environments and Behavior at UCSB is investigating the utility of using immersive virtual environment technology, including virtual humans, for social psychological research (Blascovich et al., 2002). In addition, Pertaub, Slater, & Barker (2002) conducted a study to assess the extent to which social anxiety, especially fear of public speaking, was induced by a virtual audience and the extent of influence of degree of immersion. The authors found that social anxiety was induced by the virtual audience, and that the degree of anxiety experienced was directly related to the type of feedback the speaker received from the audience. Finally, virtual humans have been proposed for use in the assessment and rehabilitation of Attention Deficit Hyperactivity Disorder (ADHD) in virtual classroom scenarios (Rizzo et al., 2000).

2 Personality Study

The purpose of this experiment was to test whether the personality of a virtual human affects the amount of time a participant spends with that virtual human. As a driving application around which to structure our research, we designed the **Di**gital **An**imated **A**vatar (Diana). Diana exhibits three personalities: *happy*, *sleepy*, and *grumpy*. *Happy* Diana has a pleasing tone of voice and facial expressions, *sleepy* Diana has a neutral tone of voice and facial expressions, and g*rumpy* Diana has an unpleasant tone of voice and facial expressions (figure 1).

Hypothesis: Diana's personality will have an effect on the amount of time participants are willing to help her on a task.

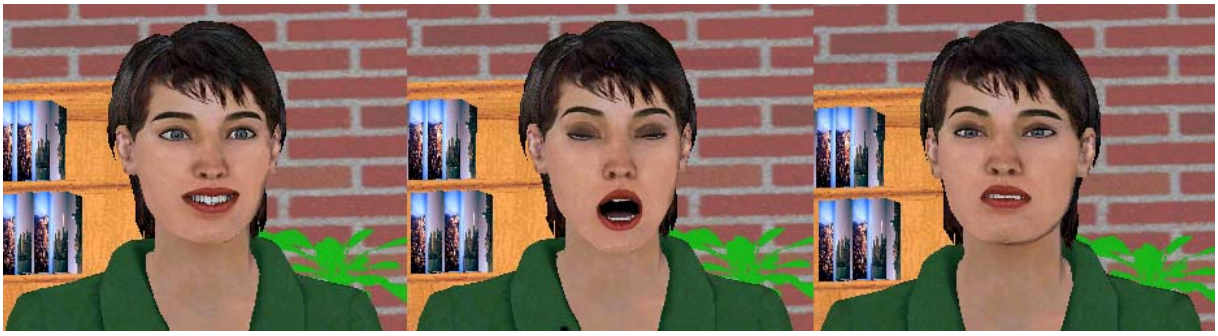

Figure 1: *Happy*, s*leepy*, and *grumpy* Diana

2.1 Participants

The participants were 17 students from the University of North Carolina at Charlotte. Thirteen males and four females between the ages of 20 and 44 participated. Volunteers were recruited from introductory computer science courses, through fliers, and by word of mouth and were not compensated for participation. Most of our participants reported that they were heavy computer users. The ethnic breakdown of the participants was as follows: nine were Caucasian, six were Asian, and two were African-American.

2.2 Task

Our intention was to create a task where the participant's role was to help Diana accomplish a goal, similar to what was done by Fogg & Nass (1997) in their study on reciprocity and computers. The particular task that we chose was to interact with Diana by answering a series of question sets about various pictures. A question set consisted of four questions about one picture. Pictures with significant items of visual interest and that were unique in content were chosen for the task. There were a total of 20 question sets. The questions were in a multiple choice form. Figure 2 shows a sample picture and the multiple choice answers to one of the questions.

Participants were instructed to "spend about ten minutes with Diana" by the experimenter. These directions were intentionally vague because we wanted the participants to respond to Diana's request for them to continue helping her. We were interested in how long participants were willing to work on a task in order to help Diana collect data on visual memory.

Participants were first taken to the experiment room and seated at a desk facing a projection screen where Diana was projected life-sized. Next, the experimenter introduced Diana and initiated the program. The experimenter then left the room. Diana re-introduced herself and asked the participant to type in his or her name. Next, Diana told the participant that he or she would be helping her with a test on visual memory. Participants interacted with Diana using a keyboard. After hearing the instructions, participants indicated when they were ready to start. Diana informed participants that she would show them a picture for five seconds and then ask them questions about the picture. A picture was then projected on the display for five seconds and automatically removed.

Next, Diana asked the participant a series of four questions about the picture. The questions were asked one at a time and the participants chose from a list of multiple choice answers presented on the display (figure 2b). The participants also had the option of repeating the question. After completing a question set, Diana told the participant his or her score. Next, Diana asked the participant to help her with another question set. If the participant chose not to help, Diana begged the participant to continue helping for a total number of three times before allowing him or her to quit. If the participants completed all the question sets, Diana informed them that the task was complete and asked them to leave the experiment room.

(a) Sample picture (b) Sample multiple choice answers

Figure 2: Sample picture and display of multiple choice answers for question: "What did the sign on the tree say?"

2.3 Apparatus

A Pentium IV 2.4 GHz Dell PC with an nVidia GeForce4 Ti 4200 graphics card rendered the virtual environment, including animating and rendering Diana. The graphics were rendered with OpenGL, and then projected using a Sony VPL-CX5 data projector. We used one of Haptek Corporation's interactive 3-D characters to build Diana.

Haptek also has a library which allowed us to create our own realistic animations and behaviors. Figure 3 shows some of the gestures Diana used to point to pictures.

In addition, Diana displays life-like behaviors such as breathing, blinking, and other subtle gestures. The Haptek library included standard moods which were used for Diana's different personalities. *Happy* Diana had a pleasing tone of voice and facial expressions, *sleepy* Diana had a neutral tone of voice and facial expressions, and *grumpy* Diana had an unpleasant tone of voice and facial expressions (figure 1). Diana interacted with the participants via pre-recorded audio. Participants' names were pre-recorded, so that *happy* and *sleepy* Diana could refer to participants by their names. *Grumpy* Diana did not use participant names. There was also a difference in the dialogue across the personalities such that *grumpy* Diana was always negative, "you did not get any wrong", *happy* Diana was always positive, "you go them all right!", and *sleepy* Diana was always neutral, "your score is 4 out of 4". Haptek's lip-sync software was used so that the audio matched Diana's lip movements.

Figure 3: Diana gesturing to pictures

2.4 Design and Procedures

The experiment was a between-subjects design. The independent variable was Diana's personality. The dependent variables were interaction time, number of question sets attempted, and co-presence scores. Participants were randomly assigned to one of three conditions, which corresponded to interacting with either *happy*, *sleepy*, or *grumpy* Diana.

Figure 4: Experimental setup

Before the experimental session, participants filled out a pre-survey. They were then taken to the experiment room where they were seated at a desk facing a screen where Diana was visually presented as a "life-sized" interactive character using a data projector (figure 4). Participants interacted with Diana using a keyboard placed on the desk. Two speakers at the bottom of the projection screen were used to output the audio.

Participants were not told how many question sets to complete, but simply to spend about ten minutes with Diana. Diana persistently asked the participant to continue on with the task after each question set. Following the interaction, participants filled out a post-survey.

3 Measures

A pre-experiment survey was used to obtain some background data about the participants. The survey included questions about participant demographic characteristics (such as age, gender and ethnicity) and it measured performance on Rotter's (1966) Locus of Control scale as well as a Computer Anxiety Scale (Cohen & Waugh, 1989).

The Computer Anxiety Scale (Cohen & Waugh, 1989) is a 16-item questionnaire that measures a participant's level of computer anxiety, and was used to screen participants in the event of an unusually high anxiety level. The Locus of Control is a 13-item questionnaire developed by Rotter (1966). It measures generalized expectancies for internal versus external control of reinforcement. People with an internal locus of control believe that their own actions determine the rewards that they obtain, while those with an external locus of control believe that their own behavior does not matter much and that rewards in life are generally outside of their control. These measures were used to help determine if there were any confounding factors affecting the results among the different conditions.

The post experiment questionnaire was used to measure co-presence, or the participants' sense of being with Diana. Co-presence was measured using the 14-item Slater Co-Presence Questionnaire (Mortensen et al., 2002). Participants used a 7-point Likert-style scale (1. *not at all* to 7. *a great deal*) to respond to a series of items (such as "I had a sense of being with Diana…." or "The experience seems to me more like interacting with a person...").

Finally, the time spent interacting with Diana and the number of question sets each participant attempted were logged.

4 Results

We collected data from 5 participants in the *happy* Diana condition, 7 in the *sleepy* Diana condition, and 5 from the *grumpy* Diana condition. The data from the computer anxiety scale, locus of control, co-presence mean and count, interaction time, and number of question sets attempted were analyzed using a series of one-way-between-subjects analysis of variance (ANOVAs) with α=0.05 level for significance.

4.1 Locus of Control and Computer Anxiety

Scores on the locus of control test were not significantly different among groups, $F<1$. In addition, the difference in computer anxiety scores were not significant among groups, $F<1$. These measures were used to help determine if there were any confounding factors affecting the results among the different conditions. There did not appear to be any significant differences among the population groups of the three conditions.

4.2 Co-presence

Figure 5 shows bar graphs and means of participants' ratings of Diana's realism, their desire to meet Diana, and their co-presence mean and count by condition.

The co-presence count (number of answers of 6 or 7 for all 14 items) was moderate overall ($M=3.06$, $SD=2.38$, $N=17$), and the difference among groups was not significant, $F(2, 14) = 1.25$, $p=0.32$. The co-presence mean was 3.60 overall ($SD=0.93$, $N=17$. The difference in co-presence mean among the groups was not significant, $F<1$.

Although participants were most interested in meeting *happy* Diana, the difference among the groups was not significant, $F<1$. Overall, participants indicated that they would like to meet Diana ($M=4.12$, $SD=2.15$, $N=17$). Participants found *happy* Diana to be the least realistic, *sleepy* Diana somewhat realistic, and *grumpy* Diana the most realistic. The difference among the groups was not significant, $F<1$. The overall realism rating was $M=3.82$ ($SD=1.24$, $N=17$).

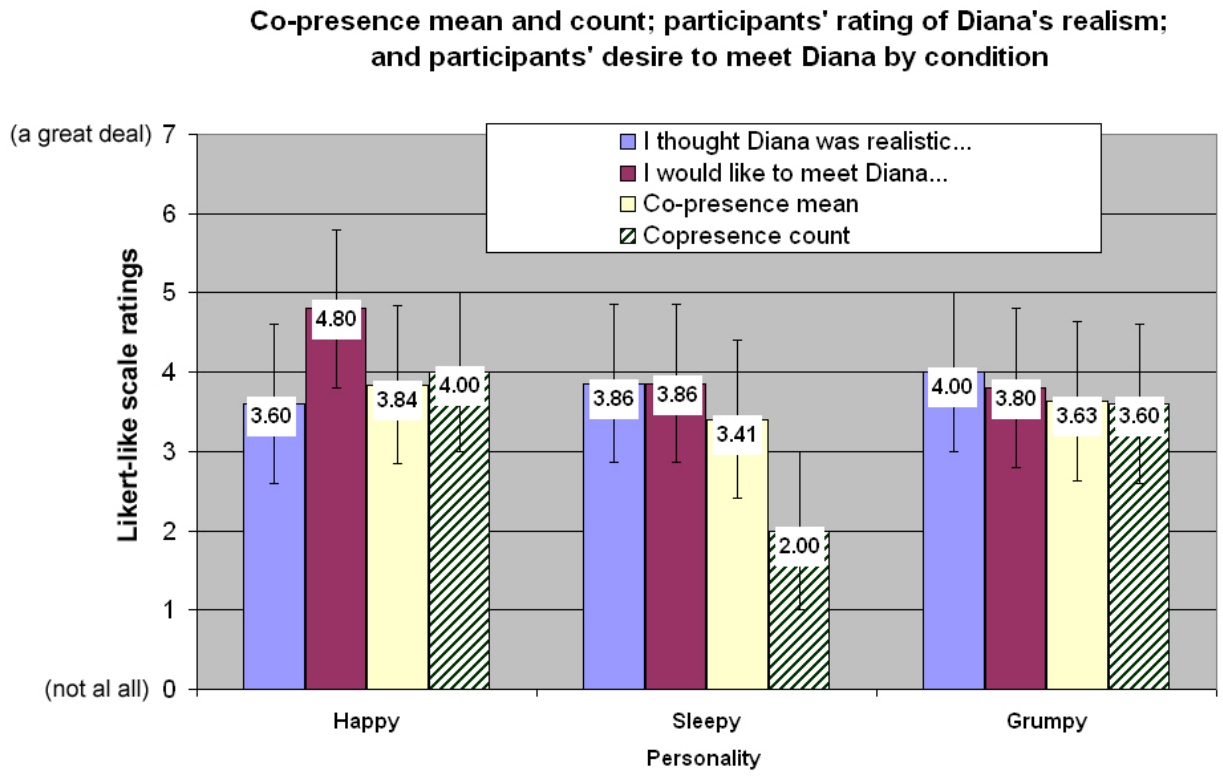

Figure 5: Bar graphs and means of participants' ratings of Diana's realism, their desire to meet Diana, and their sense of co-presence by condition

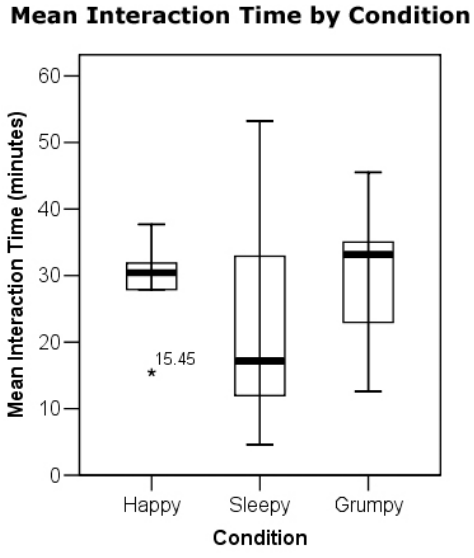

Figure 6: Boxplot of participants' interaction time with Diana by condition.

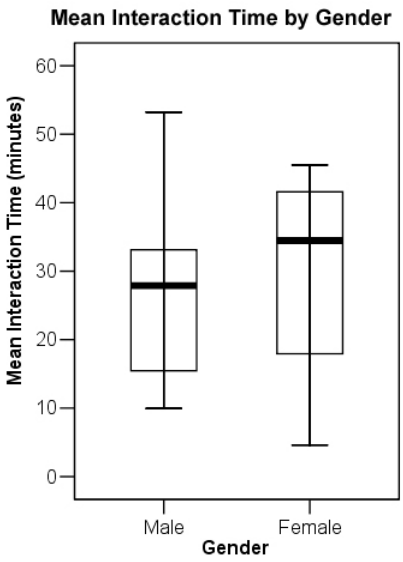

Figure 7: Boxplot of participants' interaction time with Diana by Gender.

The "box" area of the boxplot contains the 50% of the values that fall between the 25th and 75th percentiles. The heavy black line through the box is the median. The "whiskers," or lines that extend from the box, show the spread of scores from highest to lowest. Stars represent extreme values and outliers.

4.3 Interaction Time

Although it is clear that participants spent the least time with *sleepy* Diana (M=23.5, SD=17.0, N=7), the difference among the groups was not significant, F<1. Participants spent about the same amount of time with *grumpy* Diana (M=29.8, SD=12.5, N=5) and *happy* Diana (M=28.7, SD=8.2, N=5). Female participants spent more time overall with Diana (M=29.8, N=4, SD=17.8) than the male participants (M=26.0, N=13, SD=12.2). Figures 6 and 7 show boxplots of participant's interaction times with Diana by condition and gender.

4.4 Question Sets Attempted

In addition to analyzing the amount of time participants spent with Diana, we also considered the number of question sets participants attempted. We found that participants attempted the least number of question sets with *sleepy* Diana (M=10.6, SD=7.5, N=7) and the most with *grumpy* Diana (M=16.4, SD=5.7, N=5). Participants with *happy* Diana attempted 14 question sets on average (M=14.0, SD=5.6, N=5). Overall, participants attempted about 13 question sets (M=13.3, N=17, SD=6.6). The difference among the groups was not significant, $F(2, 14) = 1.21$, $p=0.33$.

It was interesting to us that participants attempted the most question sets with *grumpy* Diana. There could be many reasons for this finding. Perhaps *grumpy* Diana was the most demanding and therefore presented a more authoritative figure than *happy* or *sleepy* Diana. Also, since participants found *grumpy* Diana the most realistic (section 4.2), they may have been more willing to help her as compared to the other personalities. Finally, participants may have simply thought it was amusing that a virtual human is grumpy and having a bad day.

4.5 Participant Comments

Although there were no significant differences among groups in the amount of time participants spent with Diana, their comments reflected that they *did* distinguish her different personalities. When asked "How would you describe Diana's personality?"

Participant who interacted with *happy* Diana reported:

- "[Diana] seemed overly excited like your dog when you come home from a day at the office. She almost seemed too happy; she needed to be a slight tone down from what she is."
- "[Diana] was friendly, patient"
- "[Diana] seemed outgoing"

Participant who interacted with *sleepy* Diana reported:

- "She seemed bored with her work, I could relate, my job is so boring."
- "Bored, uninterested"
- "Tired, depressed, sleepy, caring very little about her task"
- "Bored, lacking enthusiasm"
- "Bored, she even yawned when she said my name"

Participant who interacted with g*rumpy* Diana reported:

- "[Diana] was so business-like and just wanted to get the job done!"
- "She wants to get her job done fast!!!"
- "She was very nonchalant. She seemed to not care too much about the result. When you got them all correct she did not say 'good job', but when you missed some, she'll tell you!"
- "[Diana] was unfriendly"

5 Discussion

We hypothesized that Diana's personality would have an impact on the time participants spent with her. However, participant comments suggest that this was not the case. Despite participants mentioning Diana's personality when asked what they thought of her, none indicated that her attitude was a motivating factor to complete the task or a

reason for eventually quitting. The following participant comments from the debriefing session highlight some factors that may have affected the outcome of our study:

- When asked: "What was your motivation to answer the questions?"
  - 17.6% of the participants said that they thought it was what there were *supposed to* do
  - 41.2% of the participants said that they *wanted to get more questions right*
  - 35.3% of the participants said that it was *fun to look at the pictures*
  - 35.3% of the participants said: that *Diana was begging* them to do more questions

- When asked: "Why did you eventually quit?"
  - 23.5% of the participants said that they *had to go to work*
  - 29.4% of the participants said that *Diana told them to*
  - 23.5% of the participants said that they were *not getting the questions right*
  - 29.4% of the participants said that they *felt that about 10 minutes had passed*

Some of the participants reported that their motivation for answering the questions was that "it was what they were supposed to do". The participants' desire to please the experimenter, as illustrated by their comments, may be due to what is known as demand characteristics. Orne (1962) defined demand characteristics as any features of an experiment, which help participants work out what is expected of them, and consequently lead them to behave in an artificial and unnatural way. In addition, given the cultural backgrounds of our participants (65% from the United States, and 35% from Asian backgrounds), it is valid to consider cultural difference as a possible confound. One of Hofstede's (1980) cultural dimensions, p*ower distance,* focuses on the nature of human relationships in terms of hierarchy. According to the power distance dimension, in high power distance cultures, such as Asian cultures, obedience to authority (in this case the experimenter) is expected. In low power distance cultures, such as the United States, emphasis is on equality and respecting the individual.

Another possible confound may have been the participants' desire to do well on the picture task. About 40% of the participants claimed that their motivation to complete the question sets was that they "wanted to get more questions right". This is related to what is known as the *achievement* versus *ascription* dimension of culture (Trompenaars & Hampden-Turner, 1993). In achievement-oriented cultures, such as the United States, individuals derive their status from what they have accomplished. In an ascriptive society, such as Asian cultures, individuals derive their status from birth, age, gender or wealth.

Both the entertainment factor of the pictures, and the novelty of interacting with a virtual human, may have also confounded our results. During the debriefing session, many of the participants said that "it was fun to look at the pictures". Also, many of the participants said they quit because they had to go to work. Although this is not a factor which is dependent on Diana's personality, it may have nonetheless affected the outcome of our study.

6 Lessons Learned

This study was our first attempt to assess the effectiveness of human-virtual human interaction. Our initial idea was that in real life a person's personality makes a difference in how others interact with that person, so we thought it would also effect how people interact with a virtual human. When we attempted to show this effect with a virtual human, however, we made some mistakes.

First of all, we chose time spent interacting with the virtual human as our primary measure. This measure was not sensitive enough for our study, especially considering all the possible confounding factors we discussed in the previous section.

Second, the between-subjects design may have been a poor choice for this study. A within-subjects design would have eliminated the issues outlined in the previous section, such as some of the participants' desire to do well on the task and to please the experimenter. Unfortunately, a between-subjects design does not account for such individual differences and may have further confounded our results.

In addition, we did not include a human condition where we had a real person with a *happy*, *sleepy*, or *grumpy* personality for participants to interact with. We may have gotten the same results with a real person exhibiting the same personalities. For example, many of participants were driven by the entertainment factor of the pictures in the task. Others simply wanted to please the experimenter by doing what they thought they were supposed to do.

Despite the fact that the instructions were to spend about ten minutes with Diana, the average interaction time was almost 27 minutes (M=26.9, N=17, SD=13.2)! This may have been caused by many factors, including the entertainment factor of the pictures, the desire to do well on the task, being caught up in the task, loosing track of time, or simply the enjoyment of interacting with Diana. Unfortunately, we cannot know the exact cause since we did not include a control condition where participants did the same task without interacting with a virtual human.

Finally, we did not have data from human-human interaction literature to compare our results to. This type of personality study has not been attempted in the past, so we have no ground for comparison to see if our results matched what has been done before in the field of human-human interaction.

From these lessons learned, we realized that the best way to compare human-virtual human interaction to human-human interaction was to look to social psychology research. The field of social psychology has a storehouse of experience in the study of human-human interaction, and a large amount of data from many studies to which we can compare our results. We have since started applying social psychology principles to our research with virtual humans, and have begun replicating classical tests of human-human interaction. We hope that directly comparing the results of human-human interaction with human-virtual human interaction will provide evidence of whether learning to interact with virtual humans leads to better social interactions with real humans.

7 Current and Future Work

We have already begun to replicate classical social psychology studies and replacing the human role with a virtual human. In our first such experiment, we replicated a study of the social facilitation/social inhibition effects. Social facilitation/inhibition refers to performance enhancement of a simple or well learned task, and performance impairment of a complex or novel task, when done in the presence of others. Using a virtual human, we were able to elicit social inhibition from our female participants (Zanbaka, Ulinski, Goolkasian, & Hodges, 2004).

We are currently conducting a study which replicates the persuasion effect (Petty & Cacioppo, 1986). Participants will be randomly assigned to listen to a persuasive passage delivered by a female human, a female virtual human, a female cat-like character, a male human, a male virtual human, or a male ogre-like character. We plan to compare the extent to which participants are persuaded by a virtual character as compared to a real human. We are also interested in how the gender and species of the virtual character will affect the results.

In the future, we plan to replicate more classical tests of human-human social interaction including conformity and obedience. In addition, we plan to vary the age and race of the virtual character in these future studies to examine how the changes affect the results.

References

Anderson, P., Rothbaum, B.O., & Hodges, L. F. (2003). Virtual reality exposure in the treatment of social anxiety. *Cognitive and Behavioral Practice*, 10, 240-247.

Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., & Bailenson, J. (2002). Immersive virtual environment technology: Not just another research tool for social psychology. *Psychological Inquiry*, 13, 103-124.

Cohen, B. A. & Waugh, G. W. (1989). Assessing computer anxiety. *Psychological reports*, 65, 735-738.

Fogg, B. J. & Nass, C. (1997). Do users reciprocate to computers? *Proceedings of the Computer-Human Interaction Conference*, Atlanta, GA.

Hofstede, G.H. (1980). Culture's consequences: International differences in work-related values. Newbury Park, CA: Sage Publications.

Johnsen, K., Dickerson, R., Raij, A., Lok, B., Jackson J., Shin, M., Hernandez, J., Stevens, A., Lind, D. S. (2005). "Experiences in Using Immersive Virtual Characters to Educate Medical Communication Skills," To Appear in: *Proceedings of IEEE Virtual Reality 2005*. Bonn: Germany.

Mortensen, J., Vinayagamoorthy, V., Slater, M., Steed, A., Lok, B., & Whitton, M. (2002) Collaboration in Tele-Immersive Environments. In: *Proceedings of the Eighth Eurographics Workshop on Virtual Environments (pp.* 93-101). Barcelona: Spain.

Nass, C. & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81-103.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776–783.

Pertaub, D. P., Slater, M., & Barker, C. (2002). An experiment on pubic speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments*, 11, 68-78.

Petty, R. E., & Cacioppo J. T. (1986). *Communication and Persuasion*. New York: Springer Verlag.

Rotter, J. (1966). Generalized expectancies for internal versus external control of reinforcements, *Psychological Monographs*, 80, Whole No. 609.

Rizzo, A., Buckwalter, J.G., van der Zaag, C., Neumann, U., Thiebaux, M., Chua, C., et al. (2000). Virtual environment applications in clinical neuropsychology. In: *Proceedings IEEE Virtual Reality* 2000 (pp.63-70), New Jersey: USA.

Scargle, J. D. (2000). Publication Bias: The "File-Drawer" Problem in Scientific Inference. *Journal of Scientific Exploration*, 14 (1), 91–106.

Slater, M., Howell, J., Steed, A., Pertaub, D. P., Garau, M., Springel, S. (2000). "Acting in Virtual Reality", In: *ACM Collaborative Virtual Environments 2000* (pp.103-110), San Francisco: USA.

Slater, M., & Usoh, M. (1994). Body Centred Interaction in Immersive Virtual Environments, in N. Magnenat Thalmann & D. Thalmann (Ed), *Artificial Life and Virtual Reality* (pp. 125-148). John Wiley and Sons.

Trompenaars, F. & Hampden-Turner, C. (1998). Riding the waves of culture: Understanding cultural diversity in global business (2nd ed.). New York: McGraw Hill.

Zanbaka, C., Ulinski, A., Goolkasian, P., & Hodges, L.F. (2004). Affects of virtual human presence on task performance. In: *Proceedings of the 14th International Conference on Artificial Reality and Telexistence* (pp.174-181). Coex: Korea,