ASSESSMENT OF BEHAVIORAL ENGINE TECHNOLOGIES AND SUBJECTS' RATING OF REALISM IN AN INTERVIEWER SKILLS TRAINING TOOL[1]

Michael W. Link, Ph.D., Polly P. Armsby, Laura Flicker, and Rachel Caspar

RTI International, Research Triangle Park, North Carolina, 27709-2194

KEY WORDS: technology-based training, virtual reality, refusal avoidance, speech recognition

Survey research is in an era of great challenge. Response rates across all modes of data collection have been in decline, threatening the validity and utility of the information collected in surveys. As it becomes more difficult to convince sample members to participate in surveys, it is essential that the interviewers who are on the front lines of collecting these data are given the tools they need to be successful in their jobs. Training tools built using responsive virtual human technology (RVHT) hold the promise of offering interviewers a simulated, realistic environment for developing and practicing basic interviewing skills – such as gaining respondent cooperation, probing, administering informed consent – and honing those skills over time. RVHT reduces the amount of learning that must occur on the job, by allowing repetitive practice in a virtual environment.

RVHT is admittedly in its developmental infancy and requires additional improvements before it can be deployed as a fully mature technology in a production environment. The research presented here is one small part of a larger research program shepherding the growth and development of these technologies. These analyses provide an initial assessment of the behavioral engine component (a key element of RVHT) of an RVHT tool developed to allow telephone interviewers repetitive practice in gaining respondent cooperation during the first thirty seconds of a telephone interview. The data were derived from a controlled experiment with twelve subjects, who conducted repeated tests and evaluations of the application across four different types of computer platforms (both PC and laptop-based). Using these data, we examine (1) the behavior of the application as measured by conversation exchanges, semantics, and complexity across these four computer platforms, (2) users' evaluations across the different platforms of the realism of the application in terms of response time, overall conversation, and objections raised, and finally (3) the relationship between application behavior and users' perceptions of the realism of the virtual environment. Together these analyses provide an initial base of information for better understanding, assessing, and improving RVHT-based interviewer training applications.

Background

Research has shown that flexibility is critical for developing effective interaction skills (Groves & Couper, 1998) and for performing well under time constrained, information-poor, and other difficult conditions (Klein, 1998). In order to acquire flexible and effective approaches to gaining respondent cooperation, new and experienced interviewers require a learning environment that realistically simulates the environment they face in an interviewing situation. The consistency that is gained by repetitive practice in virtual and constructive learning environments leads directly to effective decisions in the production environment (Ross, Pierce, Haltermann, & Ross, 1998). Practice also leads to increased confidence before the first real on-the-job experience, minimizing the amount of on-the-job learning that is necessary. In the survey world, on-the-job-learning can translate into numerous unsuccessful interview attempts at the start of a study by a new interviewer, leading to lower response rates, lower quality data, delayed schedules, and increased costs.

This is exactly the type of scenario in which RVHT can be most effective. The outset of any interview is generally very fluid, despite the fact that interviewers are nearly always provided with an introductory script or set of bullet points for making the introduction. Sample members often interrupt interviewers with a barrage of questions or remarks, such as "I'm not interested," "I don't have time," "How did you get this number," or "Stop interrupting our dinner time!" Non-response research suggests that the best approach to obtaining participation is for the interviewer to immediately reply with an appropriate, informative, tailored response (Camburn, Gunther-Mohr, & Lessler, 1999; Groves & Couper, 1998; Groves, 2002). Generally, such skills are taught through a combination of lecture, paired-mock practice with other interviewers, and by using multimedia to listen to real or mock audiotapes of exchanges between interviewers and sample members. RVHT allows us to take skill building to the next level, by providing a realistic, simulated environment in which an interviewer can practice and hone his or her skills.

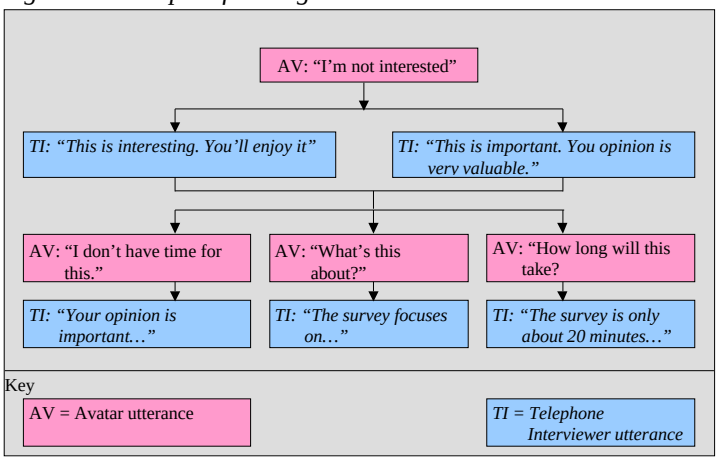
For example, in a formal telephone interview, the interviewer is generally taught to begin with a scripted introduction ("Hello, my name is … I am conducting a survey sponsored by…."). During these interactions, the trainee is expected to defuse potentially antagonistic situations by engaging in active listening, using polite, professional language and employing a calm and confident tone. If this is done successfully, the respondent will be placated and be motivated to participate.

The application tested here involves the use of an RVHT-based application to simulate the environment a telephone interviewer faces during the first thirty to sixty seconds of a telephone survey interaction. The training tool allows interviewers to practice their skills in gaining cooperation in a self-paced, realistic environment. The software is designed such that interviewers begin with an introduction and are then required to respond to a series of objections and questions raised by the "virtual respondent." The interviewer's responses are captured electronically and processed by a natural language speech processor. Based on the content of the interviewer's speech, the software launches another objection/question or

[1] This work was supported by a research grant from the National Science Foundation (Grant No. EIA-0121211). We would also like to thank Robert Hubal, Ph.D. and Curry Guinn, Ph.D. for their contributions to the development of this application and the assistance with the conduct of this experiment.

ends the conversation by either granting the interview or hanging-up the telephone (see Figure 1).

Figure 1: Example of Dialogue Flow



The application is built using an RVHT architecture called Avatalk, that enables users to engage in unscripted conversations with virtual humans (referred to hereafter as "Avatars") and hear their realistic responses (Hubal & Frank, 2001). Among the components that underlie the architecture are a Language Processor and a Behavior Engine. The Language Processor accepts spoken input and maps this input to an underlying semantic representation (where the "meaning" of the conversation exchange is interpreted), and then functions in reverse, mapping semantic representations to speech output. The application uses spoken natural language interaction (Guinn & Montoya, 1998), not text-based interaction (except for data collection during development from subject-matter experts). The Behavior Engine maps Language Processor output to virtual human behaviors. The Behavior Engine also controls the dynamic loading of contexts and knowledge for use by the Language Processor. The architecture was designed to allow the application developers flexibility in assigning general and domain-specific knowledge. Hence, the Avatars discuss relevant concerns or excuses based on specific setup variables indicating knowledge level and initial emotional state.[2]

Procedures Used to Assess the Application

The primary purpose of the experiment described here was to test the performance and behavior of the application across different hardware platforms and to obtain assessments by users to evaluate the usability, reliability, overall acceptance and overall performance of the application. Assessing gains in the persuasion skills of the test subjects was not within the scope of this assessment; rather the focus is on actual and perceived performance of the application itself. The experiment consisted of testing four hardware platforms – a Gateway PC, an IBM PC, an IBM ThinkPad, and a Dell laptop (see Table 1) – with twelve volunteer subjects. The computers were connected to color-coded headphones for easy reference and were hidden from the test subjects' sight to prevent the physical appearance of the machine from influencing their later assessment of the application's performance on that platform.

Table 1: Machines Used in Testing

Feature	Machine Number			
	1	2	3	4
Make	Gateway	IBM	IBM	Dell
Model	E-3400 SE	Intelli-station M	ThinkPad T22	Inspirion
Laptop/PC	PC	PC	Laptop	Laptop
Speed	700 MHz	2 GHz	900 MHz	1.2 GHz
RAM	256 MB	1 GB	512 MB	512 MB

Each test consisted of three separate conversations with the Avatar on the selected hardware platform. After completing those three conversations, the subject was asked to rate the experience with respect to realism of the simulation using a three-question assessment form. Subjects were to test each platform twice. In some cases, technical problems prevented the completion of a conversation, test or assessment (See Table 2 and Table 3 for test session summaries). All sessions were recorded with the subjects' permission and the tapes were later transcribed and coded for analysis.

Test subjects were recruited from staff employed by RTI International, who possessed varying levels of professional experience with telephone interviewing and supervising telephone surveys. Before each subject's session, the test administrator explained the purpose of the experiment and that the application is designed to simulate a telephone respondent. The subjects were instructed to approach the virtual interactions as they would any general telephone survey with which they had experience. They were told to assume that no household rostering or screening would be required and therefore, the Avatar was in fact the target respondent. They were further instructed to speak to the virtual respondent, answer any questions posed, respond to objections, attempt to schedule a call-back if appropriate, and generally try to prevent the Avatar from terminating the call prematurely.

Table 2: Summary of Number of Avatar/Subject Conversations Conducted - By Test Machine

Subject	Conversations by Machine Number				Total
	1	2	3	4	
1	6	6	6	0	18
2	6	6	6	0	18
3	6	6	6	6	24
4	6	6	6	0	18
5	6	6	6	6	24
6	6	6	6	3	21
7	6	6	6	6	24
8	6	6	6	6	24
9	6	6	6	6	24
10	6	6	6	6	24
11	6	3	6	6	21
12	6	6	6	6	24
Total	72	69	72	51	**264**
Mean number of conversations per Subject					22
Mean number of conversations per Machine					66

[2] For a more complete discussion of the mechanics of the application tested here, please see Link, Armsby, Hubal, & Guinn (2002).

Table 3: Summary of Number of Subject Rating Forms Completed - By Test Machine

Subject	Rating Forms by Machine Number				Total
	1	2	3	4	
1	2	2	2	0	6
2	2	2	2	0	6
3	2	2	2	2	8
4	2	2	2	0	6
5	2	2	2	2	8
6	2	2	2	1	7
7	2	2	2	2	8
8	2	2	2	2	8
9	2	2	2	2	8
10	2	2	2	2	8
11	2	1	2	2	7
12	2	2	2	2	8
Total	24	23	24	17	**88**
Mean number of rating forms per Subject					**7.3**
Mean number of rating forms per Machine					**22**

Each subject tested the four hardware platforms in random order. After completing an initial assessment of all four platforms (termed "trial one"), the process was repeated for all four machines (termed "trial two"). Conversations with the virtual respondent began with the Avatar "answering the phone," and engaging in a conversational exchange with the subject. Each conversation ended with either the virtual respondent agreeing to participate or hanging-up the telephone (indicated by a recorded dial tone). Because subjects were not allowed to see the computer screens, they could not view the visual components of the software application, including pop-up screens indicating success or failure in gaining the respondent's cooperation. Therefore, the test administrator alerted subjects when she saw the screen prompt indicating a successful or unsuccessful outcome and instructed the subject to prepare for the next trial using either the same or a new pair of headphones.

Analysis Measures

The analysis presented here comes from variables derived from two sources: (1) coded responses from the transcripts of interactions between the Avatar and the subject and (2) evaluations made by the subjects themselves.

Transcript-derived Measures

The taped conversations were first transcribed (with the transcriptions being verified by the test administrator). Then each conversation was coded to indicate unique conversation exchanges and the semantic meaning or focus of each exchange. In all, there were a total of 910 unique exchanges that were coded from the 264 conversations (which represent 88 different trials across the four machines). From the coded transcripts, three measures were developed to measure the behavior of the RVHT application:

- **Conversation Exchange:** measures the number of Avatar-subject conversational interactions. An "exchange" is defined as the pairing of an Avatar "objection" and a subject "response." Each incident of the Avatar launching an objection and the subject responding was considered a "conversation exchange." The application tested was programmed to allow a maximum of five exchanges per conversation.
- **Conversation Semantic:** measures the content or meaning of the exchange between the Avatar and the subject. Initially all exchanges were coded into one of 35 possible "semantic" categories. These 35 categories were then collapsed into six general conversation semantics: Introduction, Survey Content, Time Concerns, Selection Criteria, Survey Attributes, and Setting Callback (see Table 4 for a fuller description of these categories).
- **Conversation Complexity:** measures the number of unique semantics observed during the course of a conversation. A conversation with a larger number of unique semantics is considered to be a more "complex" conversation than one with fewer unique semantics.

These concepts are illustrated in the following example conversation:

Line	Conversation
1	Avatar: "Hello?"
2	Subject: "Hello, my name is Steve. I'm calling about a research survey."
3	Avatar: "I'm sorry, I don't have time."
4	Subject: "I understand. The survey only takes 10 minutes."
5	Avatar: "What's the survey about?"
6	Subject: "The survey focuses of healthcare issues."
7	Avatar: "I just don't have the time."
8	Subject: "Perhaps we could begin the survey and finish at a more convenient time?"

In this example, lines 1 to 8 represent a single "conversation" between the Avatar and the subject. Within this conversation there are four complete "exchanges" (exchange 1 = lines 1 & 2; exchange 2 = lines 3 & 4; exchange 3 = lines 5 & 6; and exchange 4 = lines 7 & 8). In terms of semantics, exchange 1 would be coded as "Introduction," while exchanges 2 and 4 would both be " Time Concerns" and exchange 3 would coded as "Survey Content." Because exchanges 2 and 4 involve the same semantic (Time Concerns), the complexity of this scenario would be graded as 3 (1 for Introduction + 1 for Time Concerns + 1 for Survey Content = Complexity of 3).

Subject-derived Measures

Three additional measures were developed from observations made by the subjects themselves. For each machine trial, subjects completed three separate conversations with the Avatar. After each set of conversations the subjects were asked to rate the realism of the trial in terms of responsiveness, overall conversation, and the objections raised. Each of these dimensions was rated on a seven-point scale, where 1 = not at all realistic and 7 = extremely realistic.

- **Realism of Response Times:** Did the application respond quickly enough to mirror the way in which sample members actually respond over the telephone?
- **Realism of the Overall Conversations:** Did the dialogue that took place during the three conversations generally reflect the types of dialogues (in terms of flow and content,

pace and tone) that take place with sample members at the outset of a telephone interview?

- **Realism of the Objections Raised:** Were the objections raised by the Avatar realistic and reflective of those encountered in exchanges with reluctant sample members during actual interviews?

Table 4: Description of Conversation Semantics

Semantic	Description
Introduction	Includes mentions of interviewer's name, survey introductory script language, survey organization name.
Survey Content	Includes mentions of the topics or general content of the questionnaire.
Time Concerns	Includes mentions of how long the survey will take, questionnaire length, responding to sample member complaints about lack of time.
Selection Criteria	Includes mentions of how the sample member or household was selected to participate in the survey (i.e., from inclusion on a list, randomly selected, etc.).
Survey Attributes	Includes attempts to emphasize the importance of the survey, confidential nature, that it is enjoyable.
Setting Callbacks	Includes references for the need to schedule an appointment, specifying callback days and times.

Findings

The analysis was conducted in three parts: (1) evaluation of the application's behavior across different computer platforms, (2) subject evaluations of the application's performance across four platforms, and (3) evaluation of the relationship between subjects' evaluations of realism and the behavior of the application in terms of exchanges, semantics, and complexity of the conversations.

Behavior of the Application across Platforms

Ideally, the behavior of an application should be nearly identical across any platform from which it is launched. While *performance* measures such as the speed of the application might differ from machine to machine based on characteristics such as random access memory (RAM) or machine speed (MHz/GHz), the actual *behavior* of the application should not vary greatly. The same is true of the RVHT application tested here. While the response time of the Avatalk application might be expected to vary across different computer platforms, we should expect to see few differences across *behavioral* measures, such as average number of conversation exchanges, the types of conversation semantics encountered, and the general complexity of the scenarios.

Surprisingly, this was not the case. As shown on Table 5, there was considerable variation in all three measures across the four different computer platforms tested. First, there was a statistically significant difference in the average number of conversation exchanges across the platforms. The IBM ThinkPad had a greater percentage of shorter exchanges (i.e., those ranging from 1.0 to 2.9 exchanges per conversation), while the Dell laptop had a greater percentage of longer exchanges (i.e., those in the 4.0-5.0 range). The two PCs fell between the two laptops on this measure. In sum, a greater number of exchanges were noted in tests of the Dell laptop than with any of the other three platforms.

Table 5: Behavior of the Application (Conversation Exchanges, Semantics, and Complexity) - by Computer Platform

Conversation Measures	Machine Number 1 N	1 %	2 N	2 %	3 N	3 %	4 N	4 %	Sig.[4]
Conversation Exchanges[1]									.01
1.0 – 2.9	30	41.7	21	29.2	33	47.8	9	17.6	
3.0 – 3.9	24	33.3	39	54.2	24	34.8	24	47.1	
4.0 – 5.0	18	25.0	12	16.7	12	17.4	18	35.3	
Conversation Semantics[2]									
Introduction									1.00
Yes	72	100	72	100	69	100	51	100	
No	0	0	0	0	0	0	0	0	
Survey Content									.05
Yes	24	33.3	12	16.7	21	30.4	18	35.3	
No	48	66.7	60	83.3	48	69.6	33	64.7	
Time Concerns									.01
Yes	66	91.7	66	91.7	51	73.9	42	82.4	
No	6	8.3	6	8.3	18	26.1	9	17.6	
Selection Criteria									.094
Yes	45	62.5	42	58.3	42	60.9	33	64.7	
No	27	37.5	30	41.7	27	39.1	18	35.3	
Survey Attributes									.01
Yes	33	45.8	15	20.8	27	39.1	21	41.2	
No	39	54.2	57	79.2	42	60.9	30	58.8	
Setting Callback									.01
Yes	69	95.8	57	79.2	54	78.3	42	82.4	
No	3	4.2	15	20.8	15	21.7	9	17.6	
Conversation Complexity[3]									.01
1-2 semantics	9	12.5	27	37.5	24	34.8	12	23.5	
3-4 semantics	48	66.7	39	54.2	39	56.5	33	64.7	
5-6 semantics	15	20.8	6	8.3	6	8.7	6	11.8	

[1] Conversation exchanges are computed as the number of computer-subject interactions. Range: 1-5.

[2] Conversation semantics are the elements or components of the computer-subject interaction. For this analysis six conversation semantics were coded. This table indicates the number and percentage of times these six specific semantics were a part of each computer-subject Avatalk session. "Yes" indicates the semantic was a component of the session, "No" indicates it was not a part.

[3] Conversation complexity is computed as the number of unique semantics within a computer-subject session. If a semantic occurred twice in a session, it was only coded once in the construction of this measure. Range: 1-6.

[4] Significance based on Chi-square.

Significant differences across platforms were also noted in the percentage of times four of the six conversation semantics were encountered during testing scenarios. As would be expected in an application meant to train interviewers on the first thirty seconds of a telephone survey, all of the conversations began with an introductory exchange, where the interviewer introduced themselves, the nature of the call, and the sponsor. With four of the other five semantics, however, there were significant differences noted. "Survey content" was less likely to be a focus of exchanges when tests were conducted on the IBM PC, than on the other three machines. The same pattern is noted in terms of discussion of "survey attributes," with this topic of conversation occurring less frequently on the IMB PC than on the other computers. In terms of "time concerns" these were more likely to occur using one of the two PCs, compared to the two laptops. Similarly, "setting a callback" occurred more than 95% of the time when tests were conducted using the Gateway PC, while this percentage was closer to 80% for the other three machines. Only on "selection criteria" was the behavior of the application basically identical across the four machines.

Statistically significant differences in behavior of the application were also noted in terms of the complexity of the conversations occurring within each conversation. The Gateway PC seemed to provide subjects with a more complex practice environment, when complexity is measured as the number of unique semantics encountered during a conversation. The behavior of the two IBM machines was nearly identical despite one being a PC and the other a laptop computer. The behavior of the Dell on this dimension fell between the Gateway PC and the two IMB machines.

In sum, while it was initially expected that the Avatalk application would behave similarly across the four platforms tested, what we found instead was considerable variation and no clear pattern to the variability. The behavior of the application differed significantly across the four platforms in terms of average number of conversation exchanges, the likelihood of different semantics being encountered during a testing scenario, and in the complexity of the conversations themselves. There was no clear correlation in the patterns in terms of computer make, model, speed, or memory. The source of this variability will remain a key focus of future research of this application.

Subject Rating of Application Realism by Platform

Next, we examined how the subjects themselves rated their practice experiences across the four platforms. As noted previously, the subjects each conducted three practice scenarios per machine and were then asked by the test administrator to rate on a seven point scale the realism of the response time, overall conversation, and objections raised of that experience. The identity of the platform was shielded from the subjects so the subject did not know which platform they were testing each time. Subjects completed one trial (three conversations) on each machine before repeating the process as a second trial across each machine. The results of this test are provided on Table 6.[3]

Table 6: Subject Ratings of Realism of Response Time, Overall Conversation, Objections Raised - by Machine and Test Trial

	Response Time[1]		Overall Conversation[1]		Objections Raised[1]	
	N	Mean Rating	N	Mean Rating	N	Mean Rating
Total	88	4.53	88	4.68	88	5.22
Machine Tested						
1	24	4.38	24	4.42	24	5.25
2	23	3.96	23	4.48	23	5.09
3	24	4.87	24	5.00	24	5.33
4	17	5.06	17	4.88	17	5.18
(Sig.[2])		(.096)		(.331)		(.943)
Trial						
First	45	4.40	45	4.60	45	5.24
Second	43	4.67	43	4.77	43	5.19
(Sig.[2])		(.419)		(.546)		(.733)

[1] Realism of Response Time, Overall Conversation, and Objections Raised were rated on a seven-point scale after completing three Avatalk sessions per machine. Range: 1 = Not at all realistic, 7 = Extremely realistic.

[2] Significance based on an F-test of means.

In terms of evaluating the realism of the response times across the platforms, the differences (while not statistically significant at the traditionally excepted level of $p < .05$) are suggestive of a significant difference (given the relatively small sample size of 88 and a $p < .096$ value). In terms of response time, the Dell laptop rated the highest (5.06 average rating), followed by the IBM ThinkPad (4.87), the Gateway PC (4.38), and the IBM PC (3.96). The laptops, therefore, ranked higher than the PCs in terms of subjects' ratings of their response time. Ratings were also examined across the two trials for each machine to account for any "learning" that may have occurred by subjects during the course of the testing. The two trials did not differ significantly in terms of the response time measure (trial 1 = 4.40; trial 2 = 4.67).

Subjects demonstrated little difference in their evaluations of the other two dimensions of realism – assessment of the overall conversation and the objections raised. There were no significant differences noted across each measure in terms of platform used or testing trial. Outside of possible differences in perceptions of response time, therefore, subjects found little difference in their ratings of the realism of the practice conversation generally and of the specific content of those conversations. Likewise, there appeared to be little "educating" of the subjects between trials 1 and 2.

Subject Rating of Realism Based on Application Behavior

Our third area of interest focuses on how subjects' ratings of realism may have been affected by the behavior of the

[3] Note that the unit of analysis in this section is the "trial" level (i.e., a trial equals three conversations conducted on a single machine). In all there were 88 trials conducted across the four machines For the Gateway PC and IBM ThinkPad, 24 trials were conducted. One trial had to be discarded due to a machine error in the testing of the IBM PC (leaving 23 viable trials for that platform). Finally, because of problems with the Dell laptop initially designated for this testing, a replacement was required. Consequently, only 17 trials were conducted with that machine.

application itself. We might expect that if the application is perceived to behave in a more "realistic" way that we should see differences in subjects' ratings of response time, conversation flow, and content. We found, however, surprisingly few differences in ratings on these dimensions across the different measures of Avatar behavior (conversation exchanges, semantics, and complexity).[4]

As shown on Table 7, there were no significant differences seen across these three dimensions based on the average number of exchanges per conversation within a trial. Trials with an average of 1.0 to 2.9 exchanges per conversation were not rated significantly higher or lower in terms of response time than those with 4.0 to 5.0 exchanges. The same is true when we look at ratings of the overall conversation and the objections raised during the exchanges.

Likewise, there was little variation across the three realism dimensions when we consider the six general conversation semantics. The only statistically significant difference was noted in terms of evaluation of the realism of the objections raised. When "setting a callback" was a topic of a trial, that trial tended to be rated higher in terms of the realism of the objections made, than did trials where setting a callback was not a focus.

Finally, and somewhat surprisingly, the complexity of the conversations across a trial was not related significantly to ratings of response time, nor of the realism of the overall conversation and objections raised. One might have expected that interactions that are more complex would have led to higher ratings on either or both the overall conversation or objections raised dimensions. This, however, was not the case.

Table 7: Subject Ratings of Realism of Response Time, Overall Conversation, Objections Raised - by Conversation Turns, Semantics, and Complexity

	Response Time[1]		**Overall Conversation[1]**		**Objections Raised[1]**	
	N	**Mean Rating**	**N**	**Mean Rating**	**N**	**Mean Rating**
Total	88	4.53	88	4.68	88	5.22
Conversation Turns						
1.0 – 2.9	31	4.51	31	4.68	31	5.29
3.0 – 3.9	37	4.27	37	4.49	37	5.03
4.0 – 5.0	20	5.05	20	5.05	20	5.45
(Sig.[2])		(.208)		(.294)		(.519)
Conversation Semantics						
Introduction						
Yes	88	4.53	88	4.68	88	5.22
No	0	----	0	---	0	---
(Sig.[2])		(NA)		(NA)		(NA)
Survey Content						
Yes	25	4.52	25	4.76	25	5.12
No	63	4.54	63	4.48	63	5.25
(Sig.[2])		(.958)		(.359)		(.687)
Time Concerns						
Yes	75	4.60	75	4.46	75	5.20
No	13	4.15	13	4.72	13	5.31
(Sig.[2])		(.351)		(.508)		(.799)
Selection Criteria						
Yes	54	4.51	54	4.65	54	5.11
No	34	4.56	34	4.74	34	5.38
(Sig.[2])		(.908)		(.760)		(.377)
Survey Attributes						
Yes	32	4.37	32	4.41	32	5.16
No	56	4.63	56	4.84	56	5.25
(Sig.[2])		(.479)		(.131)		(.763)
Setting Callback						
Yes	74	4.71	74	4.66	74	5.35
No	14	4.50	14	4.79	14	4.50
(Sig.[2])		(.645)		(.745)		(.035)
Conversation Complexity						
1-2 semantics	24	4.58	24	4.96	24	5.21
3-4 semantics	53	4.55	53	4.57	53	5.15
5-6 semantics	11	4.36	11	4.64	11	5.55
(Sig.[2])		(.927)		(.468)		(.699)

1 Realism of Response Time, Overall Conversation, and Objections Raised were rated on a seven-point scale after completing three Avatalk sessions per machine. Range: 1 = Not at all realistic, 7 = Extremely realistic.

2 Significance based on an F-test of means.

[4] Like the analysis presented in the previous section, the results presented here are at the "trial" level.

Conclusion

The findings presented here do not conform to what was expected going into this experiment. Rather than behaving uniformly across the four different platforms, the Avatalk application appears to have behaved differently (at a statistically significant level) on each platform with regards to the average number of exchanges per conversation, the types of resulting conversation semantics, and the semantic complexity of the conversation. At this point, there is no clear pattern or explanation for these findings. As with any software, it is important that the application does behave identically across platforms, otherwise the consistency and uniformity of the training environment could be compromised. Further research will be needed to identify and rectify these behavioral differences in the application.

In terms of user perceptions, however, there were few notable differences discerned. Subjects did not vary significantly in their evaluations of the realism of the response time, overall conversation, or objections raised across different platforms and trials for the experiment, nor across differences in the types of exchanges they encountered (shorter/longer, more/less semantically complex). In part, this may be due to the low number of observations resulting from this analysis being conducted at the trial-level (the trial level was used for analysis since that is the level at which the perception evaluations of realism were made). Further analyses will be conducted using more sophisticated statistical modeling (nested data analyses) at the conversation and exchange levels to determine if significant differences in perceptions are revealed at those levels.

Future research will also focus more closely on the other key aspect of the application – the voice recognition technologies. While not the focus of research in this particular paper, the data collected as part of this experiment did include log output data from the speech processor. These data are key to determining the behavior of the Avatar and will thus be an important component of any future analyses.

In sum, a considerable amount of basic research is still required to make RVHT applications robust, viable training tools within production environments. RVHT can hold one of the keys, however, for improved training of interviewers – both telephone and field-based staff. The research provided here offers additional information allowing developers and application designers a greater understanding of how RVHT applications respond under repeated test conditions and will hopefully help speed the development of these much needed training tools.

References

Camburn, D.P., Gunther-Mohr, C., & Lessler, J.T., (1999). Developing New Models of Interviewer
Training. International Conference on Survey Nonresponse, Portland, OR, October 28-31, 1999.

Groves, R. (2002). Principles and Practices in Non-response Reduction. Presentation at the 2002 Respondent Cooperation Workshop sponsored by the Council for Marketing and Opinion Research, New York, NY.

Groves, R., & Couper, M. (1998). Nonresponse in Household Interview Surveys. New York: John Wiley & Sons, Inc.

Guinn, C.I., & Montoya, R.J. (1998). Natural Language Processing in Virtual Reality. Modern Simulation & Training, 6, 44-45.

Hubal, R.C., & Frank, G.A. (2001). Interactive Training Applications using Responsive Virtual Human Technology. Proceedings of the Interservice/Industry Training, Simulation and Education Conference, Orlando, FL.

Klein, G. (1998). Sources of Power. Cambridge, MA: MIT Press.

Link, M., Armsby, P.P., Hubal, R., & Guinn, C.I. (2002). A Test of Responsive Virtual Human Technology as an Interviewer Skills Training Tool. Proceedings of the American Statistical Association, Survey Methodology Section [CD-ROM], Alexandria, VA: American Statistical Association.

Ross, K.G., Pierce, L.G., Haltermann, J.A., & Ross, W.A. (1998). Preparing for the Instructional Technology Gap-A Constructivist Approach. Proceedings of the Interservice/Industry Training, Simulation and Education Conference, Orlando FL.