Session 8

Capitalizing on Technology to Enhance Survey Reporting

A Comparison of the Random Digit Dialing Telephone Survey Methodology with Internet Survey Methodology as Implemented by Knowledge Networks and Harris Interactive

Jon A. Krosnick and LinChiat Chang
Ohio State University

Introduction

With their response rates declining and costs rising, telephone surveys are increasingly difficult to conduct. At the same time, Internet data collection is emerging as a viable alternative, in two forms. Some firms are distributing computer equipment to national samples recruited through RDD calling, and other firms are attracting volunteer respondents and then building panels of those individuals with some demographic characteristics distributed as they are in the nation. Most firms assemble panels of respondents who provide data on a regular basis.

Just as the survey industry was initially reluctant to embrace the telephone when it emerged decades ago as an alternative to face-to-face interviewing in respondents' homes, the field is currently uncertain about the costs and benefits of a shift to Internet-based data collection. The practical advantages of this approach are obvious: quick turn-around time, easy presentation of complex visual and audio materials to respondents, consistent delivery of questions to and collection of responses from respondents, the flexibility to allow respondents to complete questionnaires whenever they like, lack of the pressure to move quickly that is typical of telephone interviews, and the ability to track a respondent's answers across repeated waves of questioning. But potential drawbacks are obvious as well: literacy ability to read questions and navigate web pages is required, as is proficiency with a computer keyboard (and mouse when one is used); the lack of interviewers' modeling of professionalism and commitment to the task may compromise respondent attentiveness and motivation; lack of ability for an interactive conversation between a respondent and an interviewer may preclude clarifying the meanings of ambiguous questions; samples may be of uncertain representativeness, and more. Some of these potential drawbacks are overcome by internet data collection via devices other than computers (e.g., WebTV), but most remain.

Given the obvious practical advantages of Internet-based data collection, it seems worthwhile to conduct object tests of this relatively new method in direct comparison with the dominant alternative methodology: telephone interviewing. To do so, we commissioned a set of side-by-side surveys using a single questionnaire to gauge public opinion and voting intentions regarding the 2000 U.S. Presidential Election from national samples of American adults.

Data were collected by three houses: The Ohio State University Center for Survey Research (CSR), Knowledge Networks (KN), and Harris Interactive (HI). The CSR did RDD telephone interviewing. KN recruited respondents via RDD telephone interviews and equipped them with WebTV, which then permitted Internet data collection. HI respondents joined a panel after seeing and responding to invitations to participate in regular surveys; the invitation appeared on the Excite search engine web page and in various other places as well. These respondents also completed Internet surveys.

This report describes just a few of the preliminary results from our investigation. We have conducted extensive analyses of the obtained data and have much more to do analytically. The findings reported here capture a few of the general patterns we see in the data, and we look forward to providing much more extensive and detailed reports of our findings in the near future.

We compared the data from these various surveys in a number of ways:

1. We compared the demographic characteristics of the three samples to the demographic characteristics of the nation as a whole (assessed by the U.S. Census Bureau's March 2000 CPS Supplement).

2. We compared the distributions of responses to opinion and behavior questions across the three houses, expecting one of two possible patterns to be observed. If respondents answer less carefully on the Internet because of the lack of an interviewer to motivate and assist them, we thought respondents might select midpoints on rating scales more often than did telephone respondents (posited to be a form of survey satisficing; Krosnick, 1991). But if Internet respondents answer more carefully because they feel less rushed than telephone respondents do, Internet respondents might select midpoints of rating scales less often than telephone respondents. We also thought that because HI respondents were purely volunteers, their motivation to provide accurate data and therefore their response quality might exceed that of the other houses.

3. We evaluated the reliability of individual questions. If Internet respondents answer less precisely, we would expect to see higher reliability from the telephone respondents. The reverse pattern of reliabilities would indicate greater care in responding by the Internet respondents. And again, the HI respondents might have provided more reliable responses because they were volunteers.

4. We investigated the extent to which respondents manifested another form of survey satisficing: non-differentiation (i.e., identically answering a series of questions using a single rating scale). We thought this response pattern could be greater or could be less among the telephone respondents as compared to the Internet respondents, depending upon whether the Internet mode inspires more or less satisficing. If HI respondents' motivation was highest, they might have manifested the least non-differentiation.

5. Finally, we gauged the quality of responses by assessing predictive validity; stronger statistical relations between variables that theory says should be related to one another is generally taken to indicate greater respondent precision in providing the self-reports. Again, we expected that predictive validity could be either greater among the telephone respondents or less among those respondents as compared to the Internet respondents. And if HI respondents were most motivated, their predictive validity might have exceeded that of the other houses.

Data Collection

Data were collected by all three houses in two waves. The first wave of data collection was conducted before the election campaign began, in June and July. Then shortly after election day, respondents again answered questions. During the pre-election wave, respondents predicted their presidential vote and reported a wide range of attitudes and beliefs that are thought to drive vote choices. During the post-election wave, respondents reported whether they had voted and for whom they had voted.

Approximately 1,500 respondents were interviewed pre-election by telephone by the CSR. Approximately 5,000 respondents provided data to KN pre-election, and approximately 2,300 respondents provided data to HI pre-election. The CSR and HI data collections involved administering each questionnaire entirely, which lasted about 30 minutes on the telephone pre-election. KN broke the questionnaire up into three parts and administered one part per week for two consecutive weeks, took one week off, and administered the final part the next week.

Details on response rates and field periods are provided in Table 1. The pre-election response rate is highest for CSR and lower for KN. The rate at which people invited by HI to complete the pre-election survey did so is lower than the response rates for either CSR or KN. Similarly, about four-fifths of CSR and KN respondents who provided data pre-election also did so post-election, whereas this figure was 45% for HI.

Our comparisons across houses were done after weighting the samples. The weights applied to the KN and HI data were provided to us by those houses, and we generated the weights applied to the telephone data using CSR's standard procedure.

Demographic Representativeness

Table 2 shows the demographic characteristics of respondents in the CSR, KN, and HI surveys, when samples were not weighted, as well as CPS data for comparison. Under each column of percentages for a demographic variable is the average deviation of the results from the CPS figures.

In general, the average deviations are generally not huge, and sample representativeness is never dramatically poor in terms of the percentage point deviation of any survey estimate from the population. The two largest percentage point discrepancies appear between the HI and CPS percentages for people who graduated from high school and got no more education (deviation = 21 percentage points) and individuals with incomes less than $25,000 (deviation = 17.9 percentage points). Most discrepancies are much smaller than these in terms of percentage points.

The telephone survey sample manifests the smallest average deviation for three variables (education, income, and age). For two other variables (race and gender), the KN sample is more similar to the population than is either the telephone survey sample or the HI survey sample. The HI sample consistently manifests the largest average deviations from the population. As shown in the bottom row of the table, the average deviation for the telephone sample is 4.0%, 4.3% for KN, and 8.7% for HI.

Consistent with other previous studies, the telephone sample under-represents the least educated individuals and over-represents the most educated individuals. The same bias is apparent in the KN sample and even more apparent in the HI sample. Likewise, the telephone sample under-represents the lowest income individuals and over-represents higher income individuals; this bias is again more strongly apparent in the KN sample and even more apparent in the HI sample.

Again consistent with prior work, the telephone sample under-represents the youngest and oldest individuals, and these same biases are even more apparent in the KN and HI samples.

Telephone samples typically under-represent African-American respondents, and this was true here for the CSR sample, and the KN and HI samples evidenced this same bias even more strongly.

Finally, the telephone sample over-represented women, whereas the HI sample over-represented men; the KN sample's gender balance closely matched the population.

One way to summarize the discrepancies between houses is to correlate the figures in each of the first three columns of numbers in Table 2 with the numbers for the CPS in the last column. These correlations are .96 and .94 for CSR and KN, respectively, and .87 for HI. This approach again indicates nearly comparable representativeness for the CSR and KN data and less representativeness for the HI data.

Table 3 shows the distributions of the demographics after the weights have been applied to the data. As shown in the last row of the table, weighting considerably shrank the demographic deviations from the population (as should occur, of course), making the houses equivalently accurate.

Distributions of Responses

Next, we turn to examining some substantive responses to the survey questions.

Turnout. Table 4 presents post-election reports of turnout. With more than 70% of CSR and KN respondents and more than 90% of HI respondents reporting that they voted in 2000, these surveys manifest the same bias that all post-election surveys do. This may be due to self-selection: people especially interested in politics may have been especially likely to choose to participate in surveys about politics. The HI respondents also manifested the most frequent reports of having usually voted in past elections, suggesting that this sample was the most politically involved, whereas the rates for CSR and KN were quite similar.

Candidate Preference. Voters' reported choices of Presidential candidates differed between houses (see Table 4). Majorities of CSR and KN voters said they voted for Al Gore, whereas a majority of the HI voters said they voted for George W. Bush. Among non-voters, a clear plurality preferred Al Gore. Again, the CSR and KN results were quite comparable, whereas the HI non-voters manifested a more pronounced preference for candidates other than Gore and Bush.

Party Identification. The distribution of party identification confirmed two of the trends we have seen thus far (see Table 5). First, the CSR and KN data are quite similar, and the HI data are more different. Second, the HI respondents were less likely than the CSR and KN respondents to be Independents who do not lean toward either party, and the HI respondents were most likely to report strong party identification, which is again consistent with the idea that the HI respondents were the most politically involved.

Knowledge About Politics. Our pre-election questionnaire included a 5-item quiz of respondents' factual knowledge about politics, and Table 6 shows that the Internet respondents were more knowledgeable than were the telephone respondents. The average percent of questions answered correctly was 53% for CSR, 62% for KN, and 77% for HI, again suggesting the highest political involvement for the latter sample.

Other Attitudes and Beliefs. On most other measures of attitudes and beliefs, HI respondents chose the extreme ends of rating scales more often than the other respondents, while CSR respondents tended to choose the mid-points of rating scales most often. One example is displayed in Table 7, which shows the distributions of thermometer ratings of attitudes toward President Bill Clinton, Al Gore, and George W. Bush (0= least positive, 50=neutral, and 100=most positive).

Measurement Reliability

We were able to estimate the reliabilities of the measures by building a structural equation model involving two indicators of candidate preferences gathered at both waves: reported vote choice (predicted at pre-election and actual post-election) and the difference between thermometer ratings of Gore and Bush. The model posited that both measures were indicators of a latent variable (i.e., true candidate preference) at both waves, and this latent variable was allowed to manifest instability across waves. From this model, we could estimate the reliabilities of the measures (which appear in Table 8).

The CSR and KN samples yielded very comparable reliabilities, whereas the HI sample yielded notably higher reliabilities. The latter group's higher reliabilities may be attributable to more effortful reporting by those respondents and/or may be due to the HI sample containing more people who naturally answer survey questions with less random error (i.e., highly educated respondents). The structural equation modeling approach does not offer an easy way to control for demographic differences between the samples, so we cannot test these two explanations directly.

Non-Differentiation

The questionnaire included various batteries of questions using the same rating scale, and we calculated a non-differentiation score for each battery. We then standardized these scores and averaged them together to yield a single non-differentiation score for each respondent.

As shown in Table 9, the average standardized non-differentiation score was comparably high for the CSR and KN respondents and notably lower for the HI respondents. And as the regression coefficients in the first row of Table 10 show, the HI non-differentiation rate was significantly lower than those for CSR and KN, which were not significantly different from one another. This pattern remained when we controlled for differences between houses in levels of education (see row 2 of Table 10).

As the final row of Table 10 shows, though, controlling for differences between houses in terms of political knowledge revealed significantly more non-differentiation in the KN sample than in the CSR sample (b=.06, p<.05) and the HI sample (b=-.07, p<.01). Thus, the KN respondents appeared to have satisficed most according to this measure, and the HI respondents did so the least.

Predictive Validity

Finally, we examined data quality via predictive validity. These tests are all predicated on the assumption that respondents' candidate preferences should be correlated to at least some degree with the array of variables that are thought to be determinants of vote choices. We therefore conducted binary logistic regressions predicting vote choice (coded dichotomously: Bush vs. Gore) with each of its posited predictors.

These simple logistic regressions tell a consistent story: the Internet data manifest higher predictive validity than do the telephone data across the board, often substantially so. One set of illustrations of this pattern appears in Table 11. Here, the predictors are respondents' perceptions of how national conditions would change if Bush or Gore were elected President, and the dependent variable is candidate preference. The coefficients shown in columns 2 and 3 are larger than the comparable coefficients in column 1, attesting to higher predictive validity for the Internet respondents. As the first two columns of numbers in Table 12 attest, the CSR's predictive validities are consistently significantly smaller than those of KN and HI.

Note also that the predictive validity coefficients for HI (in column 3) are consistently larger than those for KN (in column 2), suggesting that HI's volunteer respondents were more precise in their reporting. As the third column of Table 12 shows, two of these five differences are statistically significant.

These differences might be attributable to differences in sample composition. That is, the KN and HI samples were higher in education and political knowledge than the CSR sample, and the HI sample was higher in education and political knowledge than the KN sample. If education and political knowledge enhance predictive validity (which they very well might), this could be responsible for the appearance of differences between the houses.

As columns 3, 4, and 5 of Table 12 show, almost all of the differences between houses are smaller when controlling for demographics and political knowledge and for interactions of the demographics and political knowledge with attitudinal predictors than when not controlling for these variables. However, all but two of the significant differences between houses remain significant after controlling for demographics and political knowledge and interactions of them with the attitudinal predictors. Therefore, the differences between houses are only slightly attributable to sample composition differences.

Specific Conclusions

These results and many others we have obtained but not reported in this memo support a set of specific conclusions:

1) Differences between the telephone and Internet samples in terms of distributions of variables or data quality were rarely huge.

2) The CSR sample was most representative of the population; the KN sample was nearly as representative; and the HI sample was least representative.

3) The Internet samples over-represented high social status individuals more than the telephone sample did, and, relative to the CSR and KN samples, the HI sample over-represented individuals highly knowledgeable about politics, individuals highly involved in politics, and individuals who voted for George W. Bush.

4) Answers given by HI respondents contained the least random error and the least systematic error attributable to survey satisficing. Rates of random error were comparable for CSR and KN, and the CSR respondents manifested the highest rates of satisficing. The differences in systematic measurement error appeared even when controlling for differences in sample composition in terms of demographics and political knowledge.

5) Reports of attitudes collected over the Internet manifested higher predictive validity than reports of attitudes collected over the telephone, and HI respondents occasionally manifested higher predictive validity than did KN respondents. The differences in predictive validity appeared even when controlling for differences in sample composition in terms of demographics and political knowledge.

General Conclusion

This study suggests that Internet-based data collection represents a viable approach to conducting representative sample surveys. Internet-based data collection compromises sample representativeness, more so when respondents volunteer rather than being recruited by RDD methods. But Internet data collection improves the accuracy of the reports respondents provide over that rendered by telephone interviews.

Reference

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology, 5, 213-236.

Table 1: Sample Sizes, Response Rates, and Field Periods

	OSU Center for Survey Research	Knowledge Networks	Harris Interactive
Pre-election Survey			
Eligible Households	3,500	7,054	12,523
Participating Respondents	1,506	4,933	2,306
Response Rate	43%	28%[a]	NA[b]
Cooperation Rate[c]	51%	31%	
Panel Completion Rate[d]		70%	18%
Start Date	June 1, 2000	June 1, 2000	July 21, 2000
Stop Date	July 19, 2000	July 28, 2000	July 31, 2000
Post-election Survey			
Eligible Households	1,506	4,143[e]	2,306
Participating Respondents	1,206	3,416	1,028
Response Rate	80%	82%	45%
Start Date	Nov 9, 2000	Nov 8, 2000	Nov 9, 2000
Stop Date	Dec 12, 2000	Nov 21, 2000	Nov 26, 2000

[a] This figure is the product of 89% (the rate at which eligible RDD-sampled telephone numbers were contacted for initial telephone interviews) and 56% (the rate at which contacted households agreed to participate in the initial telephone interview and agreed to join the KN panel) and 80% (the rate at which households that agreed to join the KN panel had the WebTV device installed in their homes) and 70% (the rate at which invited KN panel respondents participated in the survey).

[b] A response rate cannot be calculated for the HI survey, because respondents volunteered to join their panels, rather than being recruited through "cold call" contacts.

[c] This is the rate at which people who were contacted through "cold calling" and invited to participate in the CSR survey or join the KN panel ended up completing the pre-election questionnaire for this study.

[d] This is the rate at which people who had agreed to join the KN or HI panel completed the pre-election questionnaire for this study.

[e] Of the 4,933 who completed all of the first three instruments, 790 members were excluded from assignment to the follow-up survey for the following reasons: (a) temporarily inactive status (being on vacation, health problems etc.), (b) some individuals had been withdrawn from the panel, and (c) some individuals had already been assigned to other surveys for the week of the election.

Table 2: Demographic Composition of Unweighted Pre-election Samples

		OSU Center for Survey Research	Knowledge Networks	Harris Interactive	2000 CPS March Supplement
Education	Some high school	7.0%	6.7%	2.0%	16.9%
	High school grad	31.3%	24.4%	11.8%	32.8%
	Some college	19.6%	32.3%	36.6%	19.8%
	College grad	30.1%	26.0%	25.8%	23.0%
	Postgrad work	12.0%	10.6%	23.7%	7.5%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1504	4925	2306	
	Average Error	4.6%	7.4%	13.9%	
Income	<$25,000	19.0%	14.3%	12.6%	30.5%
	$25-50,000	36.9%	32.5%	32.3%	28.3%
	$50-75,000	22.0%	27.5%	25.9%	18.2%
	$75-100,000	12.9%	13.8%	14.8%	10.1%
	$100,000	9.2%	11.9%	14.5%	12.5%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1138	4335	1976	
	Average Error	6.0%	6.8%	8.6%	
Age	18-24	10.0%	7.8%	8.0%	13.2%
	25-34	17.9%	19.1%	21.2%	18.7%
	35-44	24.5%	25.8%	21.5%	22.1%
	45-54	20.7%	23.0%	27.9%	18.3%
	55-64	12.1%	12.4%	15.5%	11.6%
	65-74	9.4%	7.7%	4.8%	8.7%
	75+	5.5%	4.2%	1.0%	7.4%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1496	4923	2306	
	Average Error	1.7%	2.7%	4.6%	
Race	White	78.5%	86.4%	89.6%	83.3%
	African American	9.7%	6.9%	3.6%	11.9%
	Other	11.8%	6.7%	6.8%	4.8%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1490	4721	2183	
	Average Error	4.7%	3.3%	5.5%	
Gender	Male	45.1%	49.2%	60.1%	48.0%
	Female	54.9%	50.8%	39.9%	52.0%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1506	4910	2306	
	Average Error	2.9%	1.2%	12.1%	
TOTAL AVERAGE ERROR		4.0%	4.3%	8.7%	

Table 3: Demographic Composition of Weighted Pre-election Samples

		OSU Center for Survey Research	Knowledge Networks	Harris Interactive	2000 CPS March Supplement
Education	Some high school	17.1%	12.3%	7.9%	16.9%
	High school grad	32.7%	33.5%	36.5%	32.8%
	Some college	19.8%	28.5%	26.9%	19.8%
	College grad	21.7%	18.2%	19.8%	23.0%
	Postgrad work	8.6%	7.4%	9.0%	7.5%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1504	4925	2250	
	Average Error	0.5%	3.8%	4.9%	
Income	<$25,000	19.0%	18.0%	24.8%	30.5%
	$25-50,000	37.1%	35.3%	29.8%	28.3%
	$50-75,000	22.4%	25.8%	20.6%	18.2%
	$75-100,000	13.4%	11.9%	11.6%	10.1%
	$100,000	8.1%	9.0%	13.0%	12.5%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1138	4335	1917	
	Average Error	6.4%	6.5%	2.3%	
Age	18-24	13.5%	9.8%	14.0%	13.2%
	25-34	15.3%	19.1%	18.9%	18.7%
	35-44	22.7%	22.8%	21.8%	22.1%
	45-54	17.8%	19.8%	20.4%	18.3%
	55-64	12.4%	13.4%	10.4%	11.6%
	65-74	12.5%	9.7%	12.3%	8.7%
	75+	5.8%	5.5%	2.2%	7.4%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1496	4923	2250	
	Average Error	1.6%	1.5%	1.9%	
Race	White	83.3%	82.8%	81.1%	83.3%
	African American	11.9%	10.0%	12.3%	11.9%
	Other	4.8%	7.2%	6.6%	4.8%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1490	4721	2132	
	Average Error	0.0%	1.6%	1.5%	
Gender	Male	46.9%	49.2%	48.2%	48.0%
	Female	53.1%	50.8%	51.8%	52.0%
	TOTAL	100.0%	100.0%	100.0%	100.0%
	N	1506	4910	2250	
	Average Error	1.1%	1.2%	0.2%	
TOTAL AVERAGE ERROR		1.9%	2.9%	2.2%	

Table 4: Post-election Vote-Related Questions (Weighted Samples)

		OSU Center for Survey Research	Knowledge Networks	Harris Interactive
Usually Voted in Past Elections?	Yes	74.4%	70.2%	83.7%
	No	21.0%	22.4%	13.3%
	Ineligible	4.6%	7.4%	3.0%
	TOTAL	100.0%	100.0%	100.0%
	N	1204	3408	1028
Voted in 2000 Presidential Election?	Yes	76.5%	72.2%	90.9%
	No	23.5%	27.8%	9.1%
	TOTAL	100.0%	100.0%	100.0%
	N	1205	3406	1028
Candidate Choice of Voters	Gore	49.9%	52.5%	43.5%
	Bush	46.6%	42.9%	50.1%
	Other	3.5%	4.6%	6.3%
	TOTAL	100.0%	100.0%	100.0%
	N	881	2406	920
Candidate Preference of Non-voters	Gore	47.2%	50.2%	48.6%
	Bush	36.4%	34.1%	27.1%
	Other	16.4%	15.6%	24.3%
	TOTAL	100.0%	100.0%	100.0%
	N	253	732	91

Table 5: Party Identification (Weighted Samples)

	OSU Center for Survey Research	Knowledge Networks	Harris Interactive
Strong Republican	12.1%	12.4%	18.1%
Weak Republican	15.3%	13.5%	11.9%
Independent-Leans toward Republicans	8.6%	8.4%	8.8%
Independent-Does not Lean	23.3%	23.6%	13.6%
Independent-Leans toward Democrats	9.8%	8.7%	9.9%
Weak Democrat	17.6%	17.0%	19.0%
Strong Democrat	13.3%	16.4%	18.5%
TOTAL	100.0%	100.0%	100.0%
N	1458	4803	2250

Table 6: Percent of Correct Answers to Political Knowledge Quiz Questions (Weighted Samples)

	OSU Center for Survey Research	Knowledge Networks	Harris Interactive
Do you happen to know what job or political office is now held by Trent Lott?	21%	23%	40%
Whose responsibility is it to determine if a law is constitutional or not?	64%	78%	83%
How much of a majority is required for the U.S. Senate and House to override a presidential veto?	42%	60%	73%
Which political party currently has the most members in the House of Representatives in Washington?	64%	77%	80%
Which party would you say is more conservative?	61%	70%	73%
Average Percentage of Correct Responses per Respondent	53%	62%	77%
N	1506	4935	2250

- Average percentage of correct responses per respondent was significantly different between all pairs of houses

Table 7: Pre-election Thermometer Ratings (Weighted Samples)

Target	Rating	OSU Center for Survey Research	Knowledge Networks	Harris Interactive
President Bill Clinton	0-10	24.9%	26.9%	36.3%
	11-20	5.0%	3.6%	3.4%
	21-30	7.7%	7.7%	5.5%
	31-40	5.3%	4.3%	4.5%
	41-49	1.8%	2.3%	2.0%
	50	14.7%	11.3%	8.0%
	51-60	8.3%	6.7%	4.7%
	61-70	6.6%	5.8%	5.4%
	71-80	12.2%	14.9%	10.1%
	81-90	6.4%	8.0%	9.0%
	91-100	7.3%	8.5%	11.2%
	TOTAL	100.0%	100.0%	100.0%
	MEAN	45.4	46.5	42.6
	STD DEV	32.0	33.8	36.6
	N	1491	4698	2249
Al Gore	0-10	12.3%	18.9%	25.4%
	11-20	5.1%	4.1%	4.1%
	21-30	6.8%	8.7%	7.4%
	31-40	8.1%	7.3%	5.2%
	41-49	2.3%	3.2%	2.2%
	50	23.4%	17.1%	12.8%
	51-60	11.8%	9.2%	8.0%
	61-70	8.5%	7.0%	5.5%
	71-80	14.1%	13.9%	14.2%
	81-90	4.3%	5.7%	7.7%
	91-100	3.2%	4.9%	7.4%
	TOTAL	100.0%	100.0%	100.0%
	MEAN	49.6	47.1	46.4
	STD DEV	25.4	29.0	32.8
	N	1481	4716	2248
George W. Bush	0-10	9.6%	14.9%	18.4%
	11-20	2.3%	3.6%	4.6%
	21-30	5.9%	8.0%	8.9%
	31-40	6.5%	8.0%	5.6%
	41-49	3.3%	3.6%	3.9%
	50	20.8%	17.6%	13.5%
	51-60	13.5%	9.0%	7.1%
	61-70	10.0%	6.2%	5.6%
	71-80	19.3%	16.5%	13.7%
	81-90	5.6%	7.0%	7.1%
	91-100	3.3%	5.6%	11.6%
	TOTAL	100.0%	100.0%	100.0%
	MEAN	54.7	50.6	50.9
	STD DEV	24.4	28.4	31.7
	N	1483	4726	2249

Table 8: Reliabilities of Thermometer Ratings and Vote Choice Measures (Weighted Samples)

	OSU Center for Survey Research	Knowledge Networks	Harris Interactive
Pre-election Thermometer Rating Difference	.69	.68	.86
Pre-election Vote Choice	.94	.91	.96
Post-election Thermometer Rating Difference	.64	.65	.81
Post-election Vote Choice	.88	.88	.91
N	869	2459	910

Table 9: Average Extent of Non-Differentiation in Each House (Weighed Samples)

	OSU Center for Survey Research	Knowledge Networks	Harris Interactive
Average non-differentiation	.07 N=1478	.08 N=4847	-.05 N=2250

- CSR and KN are not significantly different from one another.
- HI is significantly different from the other two houses.
- Non-differentiation scores are standardized.

Table 10: Unstandardized Regression Coefficients Testing Differences Between Houses in the Extent of Non-Differentiation (Weighted Samples)

	Tests of Differences Between Houses			
	CSR vs. KN	CSR vs. HI	KN vs. HI	N
House Only	.01 (.03)	-.12** (.03)	-.13** (.03)	8574
Controlling for Education	.01 (.03)	-.11** (.03)	-.13** (.03)	8565
Controlling for Education and Political Knowledge	.06* (.03)	-.01 (.03)	-.07** (.03)	8565

*p<.05, **p<.01

- ♦ Standard errors are in parentheses.
- ♦ For each pair of houses (e.g., CSR vs. KN), a negative coefficient means more non-differentiation in the first listed house than the second, and a positive coefficient means more non-differentiation in the second listed house than the first.

Table 11: Effects of Expected National Conditions if Candidate is Elected (Bush - Gore) on Pre-election Vote Choice (Bush=0, Gore=1) (Weighted Samples)

	OSU Center for Survey Research	Knowledge Networks	Harris Interactive
Economy	7.19 (.48) N=1052	9.38 (.35) N=3544	9.48 (.48) N=1994
Foreign Relations	6.23 (.43) N=1056	8.35 (.31) N=3545	10.23 (.54) N=1994
Crime	5.51 (.40) N=1073	8.45 (.32) N=3548	8.78 (.45) N=1994
Race Relations	6.07 (.46) N=1069	8.41 (.34) N=3548	9.79 (.53) N=1994
Pollution	3.40 (.29) N=1064	5.76 (.22) N=3548	5.88 (.28) N=1994

- ♦ Probit coefficients appear above standard errors in parentheses.
- ♦ Expected national conditions if each candidate was elected were reported on 5-point scales ranging from “much better” to “much worse,” coded to range from 0 to 1.

Table 12: Tests of Difference Between Houses in Predictive Validity Using Pre-election Vote choice as the Dependent Variable (Weighted Samples)

	MODEL 1			MODEL 2		
Performance Domain	CSR vs. KN	CSR vs. HI	KN vs. HI	CSR vs. KN	CSR vs. HI	KN vs. HI
Economy	1.45* (.72)	1.88* (.83)	.43 (.67)	1.11 (.74)	1.06 (.86)	-.05 (.68)
Foreign Relations	1.90** (.60)	3.86* (.78)	1.95** (.68)	1.61** (.62)	3.39** (.81)	1.79* (.70)
Crime	3.12** (.55)	3.25** (.64)	.13 (.57)	2.59** (.57)	2.66** (.64)	.07 (.56)
Race Relations	2.72** (.62)	4.39** (.78)	1.67* (.70)	2.47** (.64)	3.86** (.81)	1.40* (.71)
Pollution	2.42** (.40)	2.81** (.46)	.38 (.40)	2.03** (.44)	2.48** (.51)	.45 (.44)

$^+$p<.10; * p<.05; ** p<.01

- Probit coefficients appear above standard errors in parentheses.
- MODEL 1 tests simple differences between houses.
- MODEL 2 tests differences between houses controlling for demographics and political knowledge.

Use of Responsive Virtual Human Technology to Enhance Interviewer Skills Training[27]

Michael W. Link, Ph.D., Polly P. Armsby, BA, Robert Hubal, Ph.D, and Curry I. Guinn, PhD.

Abstract
Research on survey non-response suggests that advanced communication and listening skills are among the best strategies telephone interviewers can employ for obtaining survey participation, allowing them to identify and address respondents' concerns immediately with appropriate, tailored language. Yet, training on interaction skills is typically insufficient, relying on role-playing or passive learning through lecture and videos. What is required is repetitive, structured practice in a realistic work environment.

This research examines acceptance by trainees of an application based on responsive virtual human technology (RVHT) as a tool for teaching refusal avoidance skills to telephone interviewers. The application tested here allows interviewers to practice confronting common objections offered by reluctant sample members. Trainee acceptance of the training tool as a realistic simulation of "real life" interviewing situations is the first phase in evaluating the overall effectiveness of the RVHT approach. Data were gathered from two sources -- structured debrief questionnaires administered to users of the application, and observations of users by researchers and instructors. The application was tested with a group of approximately fifty telephone interviewers of varying skill and experience levels. The research presents findings from these acceptance evaluations and discusses users' experiences with and perceived effectiveness of the virtual training tool.

Responsive Virtual Human Technology (RVHT) involves the use of natural language processing and an emotive behavioral engine to produce natural, interactive dialogues with intelligent, emotive virtual-reality (VR) agents. RVHT has great potential for use in training interaction skills, such as those required for effective survey interviewing. However, our understanding of how people interact with responsive virtual humans (a.k.a. intelligent agents) is quite limited. Better understanding requires employing RVHT in training applications and conducting systematic use, usability, perception, and training-effectiveness assessments. Important questions yet to be answered include:

- Do intelligent agents make learning more accessible?
- How willing are students to accept intelligent agents as interactive partners in learning?
- What skills can be acquired, practiced, and validated using RVHT?
- What is involved in providing a convincing simulation of human interaction, realistic enough for the student to suspend disbelief and acquire skills that will transfer to a "live" environment?

Users' interactions with RVHT applications are little studied and poorly understood. The research presented here (and the larger research program from which it is drawn) provides an initial assessment of some of the issues associated with user interface design, user acceptance of computer-based training, and perceptions of the effectiveness of the training tool. As part of this assessment, usability assessments were conducted using instructor observations and a structured questionnaire. The assessment involved the use of an RVHT-based training tool for refusal avoidance at the outset of a telephone interview. Approximately fifty telephone interviewers of

[27] This work was supported by a research grant from the National Science Foundation (Grant No. EIA-0121211) and by a Strategic Capability Development Award from RTI (R9898-002).

varying experience levels, ages, genders, races, and educational backgrounds took part in the assessment.

Background

Intelligent agents are being used in fields as diverse as computer generated (military) forces (Hill, et. al., 1998), manufacturing (Regian, Shebilske, and Monk, 1992), medicine (Miksch, Chang, and Hayes-Roth, 1996), and theater (Loyall and Bates, 1997; Lundeberg and Beskow, 1999). Intelligent agents have not been employed in training on interaction skills, although such skills are critical in a number of fields. Therefore, advanced technologies for training these "soft skills" can be a considerable asset in training. There remain, however, questions that must be answered if intelligent agents are to reach the level of sophistication required for robust interaction skills training.

Interaction skills training is certainly a new educational area in which to apply advances in information technology, such as virtual reality (VR) and agent technology. To date, VR has been shown to be effective for equipment training (Adams, 1996), maintenance training (Barnett, Helbing, Hancock, Heininger, and Perrin, 2000), simulation of military field exercises (Shlechter, Bessemer, and Kolosh, 1992), and maneuvers (Magee, 1995), and acquisition of spatial knowledge (Ragian, Shebilske, and Monk, 1992). It can be used for interaction with unobservable processes or abstract concepts (Dede, Salzman, and Loftin, 1996), tasks that are costly or dangerous to perform (Loftin and Kenney, 1994), and for gaining situation awareness (Maggart and Hubal, 1998). VR systems have become steadily smaller, faster, cheaper, and easier to use (Psotka, 1995). RTI International has integrated a spoken natural language assistant with a VR-based maintenance training environment to enhance ease of use and facilitate instruction (Guinn and Montoya, 1998). Other relevant research efforts in enabling spoken interaction with virtual humans include work done at the University of Pennsylvania (Badler, Phillips, and Webber, 1993), MIT Media Lab (Cassell and Vilhjalmsson, 1999), University of Southern California (Lindheim and Swartout, 2001), and Oregon Graduate Institute (Cole et al, 1999; Massaro et. Al, 1998).

RVHT is a relatively recent advance in training technology. Few researchers have begun integrating emotion models with agents (Becheiraz and Thalmann, 1998; Elliott, 1993; Gratch, 2000; and Klein, 1998), and none for interaction training. Portraying emotions in a virtual human, it is argued, requires clearly defined emotional states, action that shows thought processes, and accentuation to reveal feelings (Bates, 1994). In general, lifelike "pedagogical agents" can lead to improvements in problem-solving ability and can engage and motivate trainees (Johnson, Rickel, and Lester, 2000; Lester et. al, 1997). Most importantly, RVHT can open entirely new capabilities for computer-based training of interpersonal skills, and can provide the benefits of reduced training costs, individualized tutoring, and greater student convenience that are associated with computer-based training (Field, et. al., 1999).

Today, interaction skills training usually relies on peer-to-peer role playing or passive learning through videos. These approaches lead to a critical training gap, because the students are limited in the practice time and the variety of scenarios that they encounter. Nevertheless, it is exactly this practice that leads to significant on-the-job benefits.

Table 1 (adapted from Hubal, et al. 2000) presents a comparison of approaches to interaction skills training. Constraints imposed by the current approach include insufficient time in the classroom to conduct effective practice sessions, forced and unrealistic role-playing exercises, and little time or ability for individual feedback and coaching to trainees from the instructor. By using virtual humans to simulate realistic interactions, RVHT increases the amount of time trainees spend acquiring and practicing critical skills, reduces passive learning (information and skills are retained better through active learning), improves the realism of practice sessions, and enables intelligent tutoring (Graesser et al, 2000).

Table 1. Comparison of Training Approaches

Role	Traditional Approach	Role-player	RVHT Approach	Role-player
Trainee (e.g., medical practitioner, police recruit, survey interviewer)	Student's ability to learn dependent on: □ relevance of role-play scripts, □ time available during training to conduct role-plays or mock interviews, □ acting ability of role-play Partner, □ observations made by role-play Partner and/or by Instructor.	Student	Student's ability to learn enhanced by: □ using numerous age-appropriate role-play or mock interview scripts, for more practice of critical skills, □ interacting with different virtual role-play partners, □ knowing that actions are observed and tracked, □ ability to replay interaction.	Student
Conversation Partner (e.g., patient, mentally disturbed consumer, household respondent)	□ Partner must be present, available. □ Partner must act out a role that s/he will not always understand (non-essential learning activity). □ Partner is of a specific gender/age/ethnicity, limiting realism of practice.	Other person (e.g., actor, other student, Instructor)	□ Ability to simulate conditions impossible with a human. □ Standardization of responses. □ Different virtual partners of gender/age/ethnicity and having different personalities.	Virtual human
Observer/ Evaluator	□ Role-play Partner must take on second role, again a role not taken in live environment. □ Role-play Partner, if other student, is in passive learning mode.	Other person	□ Ability to track all interactions with virtual role-play partner for use in feedback, guidance, assessment. □ Knowledge of all characteristics of virtual partners.	Second virtual human
Coach/Tutor	□ Instructor must rely on role-play Partner for assessment of Student when not actually witnessing interaction. □ Only means of replaying interaction is through video, requiring an additional person and equipment.	Instructor or Supervisor	□ Virtual tutor has ability to guide learning as it occurs. □ Instructor can use automatically collected interaction information for assessment & replay, as well as actually witness interaction. □ Instructor can convey "what-if" scenarios.	Second virtual human Instructor

We stress that using virtual humans as interaction partners has disadvantages as well as advantages. Most importantly, the current state-of-the-art does not produce fully realistic conversational partners. Advances in utilizing natural language dialog features and behavior models will add tremendously to the realism. From a larger perspective, though, one must understand that virtual training is simply one component of training. Just as a trainee must "skin his/her knuckles" on actual machines in validating maintenance and diagnostic skills, so a trainee must interact with people in validating interaction skills (Helms, Hubal, Triplett, 1997). Virtual environments, though, offer advantages in reliability, repetitiveness, flexibility, throughput, and distribution that lead directly to overall cost-effectiveness of training (Field, et al, 1999).

Mechanics of the Training Application

One of the most difficult skills for a telephone interviewer to learn – and for an instructor to teach – is gaining cooperation from sample members and avoiding refusals. In telephone interviewing in particular, the first 30 seconds on the telephone with a sample member is crucial. Sample members almost automatically turn to phrases such as, “I don't do surveys,” “I don't have time,” “I'm just not interested” to avoid taking part in surveys. 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). How can the interviewer learn and then practice those responses before the survey begins, without creating more refusals during their first few weeks at work by being placed on the telephone unprepared?

The approach tested here involves the use of an RVHT-based application to simulate the environment an interview faces during the first 30 to 60 seconds of a telephone interview. The application 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 then need to respond to a series of these objections or 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 ends the conversation by either granting the interview or hanging-up the telephone (see Figure 1).

The application uses speech recognition and a behavior engine (for determining the intelligent agent’s emotional state) to produce natural dialogues with the trainees. The speech recognizer uses a basic dictionary of common words as well as a specific dictionary for each turn of a conversation. The specific dictionary consists of up to 200 words based on behavioral observations of real world events. These specific dictionaries are dynamic, therefore, changing with each turn of the conversation. During the development of the application tested here, the researchers monitored live interviews and behavior coded the responses of interviewers and sample members. These behavioral observations were then modeled, using the dictionaries and the emotional state behavior engine. Thus the specific dictionaries created for capturing responses from an interviewer to a respondent who said, “I’m too busy” in a harsh tone varied somewhat from the dictionaries created for when the respondent gave the same objection but in a softer, more reasoned tone. As trainees used the application, the emotional state of the virtual respondent varied from scenario to scenario, thus giving trainees exposure to an array of objections and emotional states. The scripts launched by the RVHT program were recorded in both a male and a female voice to add variety to the program. In all a total of six basic objections were recorded in four different tones of voice for both a male and female virtual respondent. Thus a total of 48 different practice scenarios could be offered to the trainees.

Assessment of the RVHT-based Interviewer Training Application

A primary goal of the overall research program of which this study is a part is to determine if RVHT can be an effective technology for interaction training across a broad spectrum of ethnic and socioeconomic backgrounds, jobs, and job levels. In particular, we investigate whether users find RVHT interactions accessible and acceptable. The effectiveness of this technology depends upon its ability to provide appropriate learning experiences, its ability to engage the trainee, and its acceptability to disparate users.

An "accessible" user interface is one that is easy to learn and easy to use, and can result in measurable goals such as decreased learning time and greater user satisfaction (i.e., acceptance) (Weiss, 1993). Characteristics of easy to learn and easy to use interfaces have been described as having navigational and visual consistency, clear communication between the user and application, appropriate representations, few and non-catastrophic errors, task support and feedback, and user control (Nielsen, 1993; Norman, 1993; Sneiderman, 1992; Weiss, 1993).

The assessment provided here of the interviewer training module is based on researcher / instructor observations, and user debriefings in the form of a questionnaire. Empirical data were collected on users' observed ability to interact with the application as well as their perception of the interaction. The training application was tested with a group of approximately 50 telephone interviewers of varying ages, races, experience and education levels. Trainees who participated in the assessment used the application to practice communication and thinking skills required with real conversation partners. These skills involve the use of adaptive strategies, listening and responding to the other's concerns.

To evaluate the *accessibility* of the application we focused on the following:

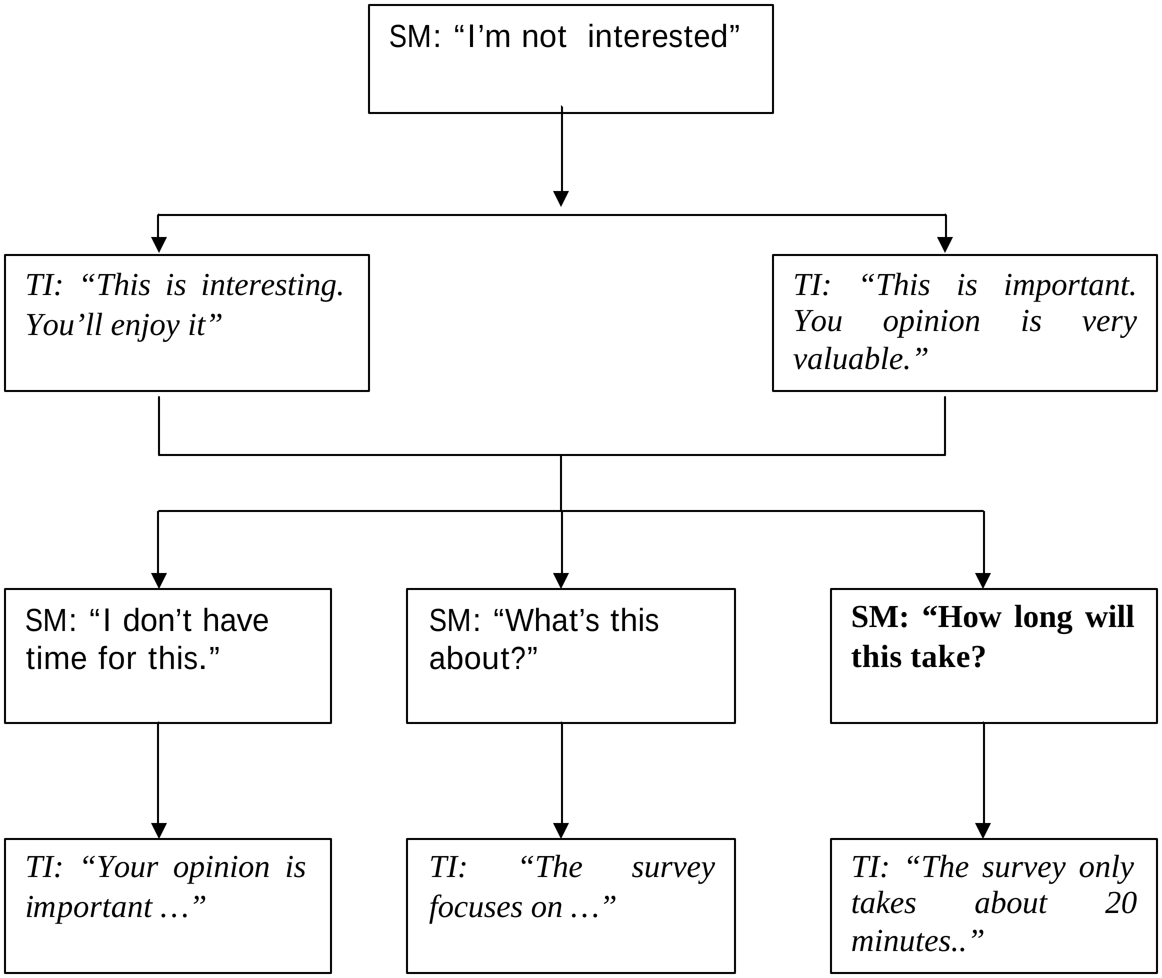
- Do users understand the basic features of the application?
- Are users able to complete each task and exit the application?
- Do users understand where they are in the application?
- Are different users (e.g., based on age, time on the job, and education level) equally able to use the application?

Instructor/researcher observation was used to assess more directly the interaction between the user and the training application, addressing questions such as:

- When there are problems (e.g., the virtual human seems to respond inappropriately), what are user reactions?
- Are inappropriate responses due to a programming error, misunderstanding in the interaction, or incorrect user behavior?
- What knowledge engineering improvements will lead to better recovery by the application when inappropriate responses occur?

Analysis of these questions will provide clues as to how smoothly the application runs, or when and why difficulties arise in its use.

Figure 1
Example of Dialogue Flow


SM: “I’m not interested”
TI: “This is interesting. You’ll enjoy it”
TI: “This is important. You opinion is very valuable.”
SM: “I don’t have time for this.”
SM: “What’s this about?”
SM: “How long will this take?
TI: “Your opinion is important …”
TI: “The survey focuses on …”
TI: “The survey only takes about 20 minutes..”

The question of whether and why participants "accept" or "reject" the virtual training environment is also central to this research. To evaluate *acceptance* of the application by the trainees, we debriefed participants using a structured questionnaire to gauge reactions and engagement in the application. In particular we are interested in the following:

- Are the virtual humans realistic enough for the users? Why or why not?
- How fast and accurate is the speech recognition?
- When recognition is inaccurate, does the application respond reasonably?
- Overall, do the users "buy into" the virtual environment?
- Could trainees detect changes in the emotive states of the virtual human using only audio cues?
- Did the trainee perceive any gains in skills from using the application?
- Would they use the application again and/or recommend it's use by others?

While some of these acceptance measures may be particular to the specific application tested, most help in gaining a general understanding of user satisfaction and affect with RVHT.

As part of the evaluation process, data were collected using a questionnaire filled out by the interviewers and notes made by instructors and researchers who observed the training sessions. The questionnaire asked questions related to users' perceptions of the realism of the interactions with the "virtual human," ease of use of the software, the perceived effectiveness of the training sessions, and some basic background characteristics of the users. In all, a diverse group of 48 interviewers filled-out the questionnaires (96% of the software users). A breakdown of some of the demographic characteristics of this set of users is provided on Table 2. Finally, each training session was observed by either the researchers or training instructors, who made notes of their observations. These observations are included as part of the analysis.

Findings

The questions posed to the interviewers were designed to assess their perceptions and experiences in using the RVHT training tool in four basic areas: ease of use of the software, realism of the training environment, impact on skill development, and desire to recommend or use the software again. Although this is the first detailed look at how users interact emotive intelligent agents for soft-skills development, we can formulate some hypotheses regarding how different types of users might respond based on how users generally differ in their use and acceptance of other computer-based tools. For example, we might expect to find that trainees who are younger, have more education, and are more comfortable using computers in general to have fewer difficulties in using the system. Likewise, we might expect that more experienced interviewers might not find the training tool as useful as inexperienced interviewers because the more experienced interviewers will have already developed and honed their refusal avoidance skills (a supposition that mirrors the finding of Groves, 2002). To examine possible differences in accessibility and acceptance of the program, we cross-tabulated all of the closed-ended questions in the questionnaire with the demographic variables listed on Table 2. Significant differences are noted in the text.[28]

[28] Because of the small number of observations (N=48) we also created dichotomous variables for both the dependent variables (collapsing scales where possible) and independent variables (collapsing or combining variables with 3 or more values). These variables were also examined to determine if

Table 2
Demographics of RVHT Trainees

Characteristic	N	%
Sex		
Male	12	25%
Female	36	75%
Education		
High School/GED	2	4%
Some College	12	25%
Four Year Degree	25	52%
Advanced Degree	9	19%
Age		
18-21	7	15%
22-29	17	35%
30-39	8	17%
40-49	7	15%
50+	9	18%
Race		
African-American	34	70%
White	7	15%
Hispanic	7	15%
Experience		
< 1,000 hours	19	40%
1,000 – 1,999 hours	17	35%
2,000+ hours	12	25%
Comfort with Keyboard		
Slow-touch typing	15	31%
Fast-touch typing	33	69%

significant differences among subgroups could be identified. Significance was evaluated at the ($p < .10$) level.

Table 3
Interviewer's Evaluation of the RVHT Training Software

	Extremely	Very	Somewhat	Not Too	Not At All
In general, how easy was the application to use?	52.1% (25)	31.3% (15)	12.5% (6)	4.2% (2)	0 % (0)
In general, how realistic did you find the overall conversation with the "virtual respondent"?	2.1% (1)	14.6% (7)	43.8% (21)	16.7% (8)	22.9% (11)
In general, how realistic did you find the objections, concerns, questions posed by the "virtual respondent"?	12.5% (6)	35.4% (17)	39.6% (19)	8.3% (4)	4.2% (2)
How easily could you determine the "virtual respondent's" emotional state or attitude based on the tone of his/her voice?	22.9% (11)	43.8% (21)	29.2% (14)	4.2% (2)	0% (0)
How easily could you determine the "virtual respondent's" emotional state or attitude based on the words used or objectives raised by him/her?	8.3% (4)	54.2% (26)	27.1% (13)	10.4 % (5)	0% (0)

Ease of Use of the Application

Training software should be accessible to users; that is, it should be relatively easy to use. As shown on Table 3, users of the RVHT software seemed to find it very accessible to use, with 84% indicating the software was either extremely easy or very easy to use (52% extremely, 31% very, 13% somewhat, 4% not too, 0% not at all). Nearly everyone found the written instructions (96%) and the verbal instructions (98%) that accompanied the training to be clear and accurate. Only eight (17%) of the 48 trainees indicated that they required additional assistance to use the training software (after the initial training received by all trainees).

The only significant difficulty encountered by the users were "insufficient memory" errors received on some of the training stations. The version of the application tested did, at times, use up considerable CPU memory. Once the machines were adjusted to handle the software memory requirements, the error messages were no longer an issue.

Realism of the Training Environment

The promise of RVHT-based training tools is that they can simulate a "real" environment, thereby allowing trainees repetitive practice in conditions that are as close as possible to what they will encounter on the job. For this particular application, the "virtual respondent" needed to mirror the behaviors and emotions of real respondents encountered when doing live interviewing. This means delivering an array of objections to the trainees in different tones of speech and emotional levels in a fast-paced manner. Interviewers were asked a series of questions to try to assess how well they accepted the virtual environment as a substitute for real work conditions. In other words, do they "buy-into" the virtual environment?

The answer is somewhat mixed. In general, trainees did not find the virtual environment to be realistic and they cited two primary reasons: the slowness of the response of the "virtual respondent" and the limited number of different objections/questions offered by the "virtual respondent." They did, however, find the responses that were offered to be realistic and stated that they could detect and respond to changes in tone and emotional cues offered by the "virtual respondents." A majority of the trainees also indicated that they felt the sessions helped them to improve their skills needed at the outset of an interview either somewhat or a lot.

When asked, *In general, how realistic did you find the overall conversation with the 'virtual respondent*,' 17% said they thought it was extremely or very realistic, 44% said it was somewhat realistic, 17% not too realistic and 23% not at all realistic (see Table 3). Slowness of the "virtual respondents" in replying (due to the lag caused by the speech recognizer as it interpreted the interviewer's responses and determined the next script to launch) was the primary problem cited by interviewers. Over three-quarters (77%) of the users felt the response time was too slow (4% felt it was too fast and 19% indicated the speed was just right). Perhaps not surprisingly, trainees who describe themselves as "fast-touch typists" were more likely than those who indicated they were "slow-touch typists" to say the response time was too slow (82% fast-touch vs 67% slow-touch; p < .08 chi-sq). Interviewers who are more comfortable at a keyboard and who, it can be surmised, tend to get through an interview faster were the ones most put-off by the perceived slowness of the response time.

The trainees were, however, more positive when evaluating the realism of the objections and questions offered by the "virtual respondent." A plurality (48%) indicated that the content of what was said was either extremely or very realistic, with 40% saying it was somewhat realistic, 8% not too realistic, and 4% not at all realistic. They also felt it was relatively easy to determine the emotional state of the virtual respondent based on the tone of voice they heard (23% extremely easy, 44% very easy, 29% somewhat easy, and 4% not too easy; no one indicated that they could not determine the avatar's emotional state from the tone of the "virtual human's" voice). Likewise, the content of the speech used by the avatar was also a good cue to trainees as to the "virtual human's" emotional state: 8% extremely easy to tell, 54% very easy, 27% somewhat easy, 10% not too easy, 0% not at all easy.

Being able to recognize changes in the emotional state of the virtual respondent changed – at least in the minds of many trainees – how the interviewer approached the situation. Nearly 60% indicated that they behaved differently in the practice scenario based on the tone of the virtual respondent's voice. Interestingly, a higher percentage of women than men reported reacting differently to the changing tone of the avatar's voice (women 67% v. men 33%, p < .04 chi-sq.). Similarly, 54% said they treated the situation differently based on the actual words used by the avatar in expressing a concern or voicing an objection. There were, however, no differences between men and women on this question. When asked how they behaved differently, interviews said they tended to soften and take a more conciliatory tone when the virtual respondent seem to grow more hostile or angered, and they mirrored the tone when the virtual respondent seemed more pleasant. Likewise, they reported tailoring the content of their responses to try to meet the objections or questions of the virtual sample member rather than simply moving forward with their script. It seems, therefore, that the both the content of the objections raised by the virtual respondent and the emotional behavior of the "virtual human" were generally accepted by the trainees and caused them to react differently within the various training scenarios.

When asked in an open-ended format to list some of the problems with the realism of the software, many cited the slowness and others indicated that the limited number of objections raised by the virtual respondent made the sessions less realistic than what they encounter on the telephone. Because this was the first iteration of the software, a conscious decision was made at the design phase to maintain a limited set of six main objections and questions ("I'm not interested," "I'm too busy," "What is the survey about?", "I don't have time right now," "How was I selected?", and "How long will this take?"). These six responses, however, were recorded in four different tones of voice (ranging from calm to upset) and recorded in both a male and a female voice. A total of 48 possible practice scenarios were, therefore, actually possible (6 responses * 4 tones of voice * 2 sexes). It appears, however, that while the interviewers do recognize and react to the different emotional cues they obtain from the different scenarios, they don't necessarily process these as being very distinct. They focus more on the actual content of the argument (regardless of the tone of voice or whether the voice is a male or female) when considering how diverse the scenarios offered are. In designing future versions of the software this will need to be considered to increase interviewer acceptance of the training tool as a realistic simulation of the environment within which they must work.

Impact on Skill Development

The purpose for allowing trainees to operate within a virtual environment is to allow them to develop and hone essential skills before entering the "real" environment, thereby reducing the amount of "on the job" skill development required. For telephone interviewers, this means an opportunity to practice their skills at gaining cooperation at the outset of an interview. Practice in a virtual environment, it is hoped, will allow interviewers – particularly new interviewers – to develop, practice, and hone these skills before getting on the telephone. New interviewers can do considerable damage at the outset of a telephone study, generating a large number of refusals as they gain comfort and confidence on the telephone. If practice within a virtual environment at the beginning of a project can reduce the numbers of initial refusals even modestly, then the training program will have value. While longer-term assessments of the effectiveness of the RVHT software will need to include examination of more objective measures of improved performance, this preliminary assessment focused on the user's assessment of the impact of the training on their own skill development.

Trainees were asked to evaluate if they thought the RVHT software increased their abilities in six different areas (see Table 4). Nearly three-quarters of the trainees felt that the practice sessions increased a lot or somewhat their ability to respond to questions and concerns by sample members. Approximately 56% felt it helped them a lot or somewhat in better gaining respondent cooperation at the outset of an interview. Likewise, over half felt it helped in their ability to adapt to differences in respondents' tone or voice or perceived moods and to adapt to differences in the speed and pace of different sample members' speech. About half of the trainees also thought that the sessions helped them a lot or somewhat in avoiding refusals at the outset of an interview.

Table 4
Interviewer's Perceptions of Effectiveness of RVHT Training Software

	A Lot	Somewhat	A Little	Not at All
Respond to questions / concerns raised by sample members	25.0% (12)	47.9% (23)	16.7% (8)	10.4% (5)
Better gain respondent cooperation during the first seconds of a call	25.0% (12)	31.3% (15)	29.2% (14)	14.6% (7)
Enhance your ability to adapt to differences in respondents' tone/mood	25.0% (12)	29.2% (14)	29.2% (14)	16.7% (8)
Think on your feet	20.8% (10)	39.6% (19)	27.1% (13)	12.5% (6)
Enhance your ability to adapt to differences in respondents pace of speaking	18.8% (9)	33.3% (16)	27.1% (13)	20.8% (10)
Avoid refusals at the outset of an interview	16.7% (8)	35.4% (17)	31.3% (15)	16.7% (8)

Once again, while more objective measures of increased ability to gain cooperation from sample members are needed in the longer-term evaluation of this training tool, it does appear that trainees perceive an increase in their ability to deal with various facets of the opening of an interview as a result of their training sessions.

Would They Use The RVHT Training Tool Again?

An effective training tool is also one that trainees should enjoy using, would use again, and recommend to others (see Table 5). Approximately two-thirds (65%) of the users said that they found using the RVHT software to be fun and enjoyable. Interestingly men were significantly more likely than women to say that they found the sessions to be enjoyable (92% men vs. 56% women, p < .05 chi-sq). Nearly three-quarters (73%) said they would like to use the software again. In addition, 83% said they would recommend the program as a training tool for other interviewers. In open-ended responses, a number of interviewers indicated that it would be a very good practice vehicle for new or less experienced interviewers.

Conclusions

This initial assessment of an RVHT-based training tool for telephone interviewers provides some valuable insights into how trainees access and accept virtual environments as practice labs and "virtual humans" as training partners. There were aspects of the training program that interviewers clearly liked, such as the ability to do repeated practice of frequently asked questions, being able to distinguish different emotional states from the tone of voice and speech content of the virtual respondent, and the opportunity to learn to think on their feet in a simulated environment before being placed into a live interviewing situation.

Table 5
Recommendation for Future Use of RVHT Training Tool

Assessment Questions:	**Yes**	**No**
Would you recommend the RVHT program as a training tool for other interviewers?	83% (40)	17% (8)
Would you like to use the RVHT program again as a training tool?	73% (35)	27% (13)
Was using RVHT fun and enjoyable?	65% (31)	35% (17)

There were also aspects that the interviewers did not like, such as the slowness of the response of the virtual respondent and the perceived lack of variety in the scenarios that were presented. This provides constructive feedback for the engineering and improvement of the software. Adding additional scenarios is a relatively easy process, involving research into the "normal" flow of such scenarios and simple scripting and programming. The responsiveness issue is a more fundamental matter, reflecting the current state-of-the-art in speech recognition. For virtual training partners to be more readily accepted, the underlying speech recognition technology needs to be improved, providing quicker, more efficient processing of the input from interviewers and more rapid launching of responses by the virtual respondent. While our research focused on a specific training application, the results have implications for a broader range of training and educational RVHT-based tools. The lessons learned here can be used to inform the development of tools in these other areas.

We do not anticipate RVHT-based training will replace instructor-led training, but we expect that combinations of RVHT-based training and instructor-led training will significantly reduce training development costs (with new development tools) and training delivery costs, while increasing trainee throughput and maintaining training effectiveness and consistency. As an additional return-on-investment, RVHT-based training can provide inexpensive, focused sustainment (i.e., refresher) training. We feel it is important to continue to investigate more robust and effective RVHT models and more efficient means of creating the models, to better understand user preferences and acceptance of RVHT, and to determine how best to use RVHT in combination with other training methods to provide cost-effective training on critical interaction skills.

References

Adams, N. (1996). A Study of the Effectiveness of Using Virtual Reality to Orient Line Workers in a Manufacturing Environment. Motorola University, unpublished dissertation.

Badler, N.I., Phillips, C.B., & Webber, B.L. (1993). Simulating Humans: Computer Graphics, Animation, and Control. Oxford Univ. Press.

Barnett, B., Helbing, K., Hancock, G., Heininger, R., & Perrin, B. (2000). An Evaluation of the Training Effectiveness of Virtual Environments. Presented at the Interservice/Industry Training, Simulation and Education Conference. November 30, 2000, Orlando, FL.

Bates, J. (1994). The Role of Emotion in Believable Agents. Communications of the ACM, Special Issue on Agents, July, 1994.

Becheiraz, P., & Thalmann, D. (1998). A Behavioral Animation System for Autonomous Actors personified by Emotions, Proceedings of the First Workshop on Embodied Conversational Characters (WECC '98), Lake Tahoe, California.

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.

Cassell, J., & Vilhjálmsson, H.H. (1999). Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous. Autonomous Agents and Multi-Agent Systems: 2, 45-64.

Cole, R., et. al. (1999). New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. In Proceedings of ESCA-MATISSE ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education, London, UK.

Dede, C., Salzman, M., & Loftin, B. (1996) ScienceSpace: Research on Using Virtual Reality to Enhance Science Education. In P. Carlson & F. Makedon (Eds.), Proceedings of the 1996 ED MEDIA Conference (pp. 172-177). Charlottesville, VA: Association for the Advancement of Computers in Education.

Elliott, C. (1993). Using the Affective Reasoner to Support Social Simulations. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, pp. 194-200, Chambery, France, August 1993. Morgan Kaufmann.

Field, S.S., Frank, G.A., Helms, R.F., & Hubal, R.C. (1999). Army Learning & Training Effectiveness Symposium. Final Report, March 30, 1999. Submitted to Battelle RTP Office, Subcontract # DAAH04-96-C-0086, Agreement #99020, Delivery Order #0366, Dated February 12, 1999.

Graesser, A., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the Tutoring Research Group (2000). AutoTutor: A simulation of a human tutor. Journal of Cognitive Systems Research, 1,35-51.

Gratch, J. (2000). Modeling the Interplay Between Emotion and Decision-Making. Proceedings of the Ninth Conference on Computer Generated Forces & Behavioral Representation, May 16-18, 2000, Orlando, FL.

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 and Training, pp. 44-55.

Helms, R.F., Hubal, R.C., & Triplett, S.E. (1997). Evaluation of the Conduct of Individual Maintenance Training in Live, Virtual, and Constructive (LVC) Training Environments and their Effectiveness in a Single Program of Instruction. Final Report, September 30, 1997. Submitted to Battelle RTP Office, Subcontract # TCN 97031, Delivery Order #0027, Dated April 16, 1997.

Hill, R., Chen, J., Gratch, J., Rosenbloom, P., & Tambe, M. (1998). Soar-RWA: Planning, Teamwork, and Intelligent Behavior for Synthetic Rotary Wing Aircraft, Proceedings of the Seventh Conference on Computer Generated Forces & Behavioral Representation, May 12-14, 2000, Orlando, FL.

Hubal, R.C., Kizakevich, P.N., Guinn, C.I., Merino, K.D., & West, S.L. (2000). The Virtual Standardized Patient-Simulated Patient-Practitioner Dialogue for Patient Interview Training. In J.D. Westwood, H.M. Hoffman, G.T. Mogel, R.A. Robb, & D. Stredney (Eds.), Envisioning Healing: Interactive Technology and the Patient-Practitioner Dialogue, 133-138. IOS Press: Amsterdam.

Johnson, W.L., Rickel, J.W., & Lester, J.C. (2000). Animated pedagogical agents: face-to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education, 11.

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

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B., & Bhogal, R. (1997). The Persona Effect: Affective Impact of Animated Pedagogical Agents. In S. Pemberton (Ed.), Human Factors in Computing Systems: CHI'97 Conference Proceedings. 359-366. New York: ACM Press.

Lindheim, R., & Swartout, W. (2001). Forging a New Simulation Technology at the ICT. Computer, 34(1), 72-79.

Loftin, R.B., & Kenney, P.J. (1994). Virtual Environments in Training: NASA's Hubble Space Telescope Mission. 16th Interservice/Industry Training Systems & Education Conference, Orlando, FL.

Loyall, A.B., & Bates, J. (1997). Personality Rich Believable Agents That Use Language. Proceedings of the First International Conference on Autonomous Agents, February 1997, Marina del Rey, CA.

Lundeberg, M., & Beskow, J. (1999). Developing a 3D-Agent for the August Dialogue System. Proceedings from AVSP '99, Santa Cruz, CA.

Magee, L.E. (1995). Virtual Reality and Distributed Interactive Simulation for Training Ship Formation Manoeuvres. Proceedings of the 36th NATO Defense Research Group (DRG) Seminar.

Maggart, L.E., & Hubal, R.C. (1998). A Situation Awareness Model. In S.E. Graham & M.D. Matthews (Eds.), Infantry Situation Awareness: Papers from the 1998 Infantry Situation Awareness Workshop. U.S. Army Research Institute: Alexandria, VA.

Massaro, D.W., Cohen, M.M., Beskow, J., Daniel, S., & Cole, R.A. (1998). Developing and Evaluating Conversational Agents. In Proceedings of Workshop on Embodied Conversation Characters (WECC), Lake Tahoe.

Miksch, S., Chang, K., & Hayes-Roth, B. (1996). An Intelligent Assistant for Patient Health Care. Knowledge Systems Laboratory Report No. 96-19, Stanford University.

Mills, K.C., Parkman, K.M., Smith, G.A., & Rosendahl, F. (1999). Prediction of driving performance through computerized testing: High-risk driver assessment and training. Transportation Research Record, 1689, 18-24.

Nielsen, J. (1993). Usability Engineering. Boston: Academic Press.

Norman, D.A. (1993). Things That Make Us Smart . Reading, MA: Addison-Wesley.

Psotka, J. (1995). Immersive Training Systems: Virtual Reality and Education and Training. Instructional Science 23, 405-31.

Regian, J.W., Shebilske, W.L., & Monk, J.M. (1992). Virtual reality: An instructional medium for visual-spatial tasks. Journal of Communication 42, 136-149.

Rousseau, D., & Hayes-Roth, B. (1997). Improvisational Synthetic Actors with Flexible Personalities. KSL Report #97-10, Stanford University.

Shlechter, T.M., Bessemer, D.W., & Kolosh, K.P. (1992). Computer-Based Simulation Systems and Role-playing: An Effective Combination for Fostering Conditional Knowledge. Journal of Computer Based Instruction 19, 110-114.

Schneiderman, B. (1992). Designing the User Interface: Strategies for Effective Human-Computer Interaction. Reading, MA: Addison-Wesley.

Weiss, E. (1993). Making Computers People-Literate. San Francisco: Jossey-Bass.

Discussion Paper: Capitalizing on Technology to Enhance Survey Response

Carol C. House
National Agricultural Statistics Service

This session of our workshop focuses on "Capitalizing on Technology to Enhance Survey Response". Before commenting on the two papers presented in the session, I will speak more broadly about this topic. First I ask: what do we mean by "technology" in this context of survey response. There exists, in fact, a fairly broad smorgasbord of related technologies, each providing tools useful in the survey process. It may be helpful for discussion purposes to group these technologies into the following categories:

- Tools for Use by Interviewers
- Tools for Use by Respondents
- Tools for Use by the Survey Designer or Survey Administrator

These are not mutually exclusive categories, and a given technology may be an excellent tool in more than one area.

Moving to the concept of "enhancing survey response", we can fashion a similar taxonomy. Clearly, as survey practitioners we want both more response (i.e. higher response rates) and improved quality for the responses that we do receive. In order to increase responses rates, we generally need either to make it easier for potential respondents to respond, i.e. reduce the burden of responding, or we need to provide more compelling reasons why they should spend their time and give up information. Thus, technology may be used to enhance survey response in any of the following ways:

- Make it easier to respond
- Communicate compelling reasons to respond
  - Facilitate quality responses

It is informative to look at which tools enhance response in which ways. Figure 1 provides a matrix of technologies, with columns defined by the entity who uses the technology and rows defined by how the technology enhances response. This is not intended to be a complete listing of technologies, and I may have left out some important examples. However, the clustering of technologies provides some insight.

Figure 1: Technology tools to enhance survey response, categorized by who uses the tool and the purpose of the tool.

PURPOSE OF TOOLS	**TOOLS FOR USE BY**		
	INTERVIEWER	**RESPONDENT**	**DESIGNER OR SURVEY ADMIN.**
EASE BURDEN OF RESPONSE	CATI / CAPI Wearable technology Geo-positioning devices	CASI Touch-tone / Voice Recognition Fax Email Web Collection Other EDR Methods	**CATI / CAPI**
COMMUNICATE REASONS TO RESPOND			RVNT – Training
IMPROVE RESPONSE QUALITY	CATI / CAPI Wearable technology Geo-positioning devices	Web Collection CASI CAPI	CATI / CAPI RDD Methods OCR / Scanning

Early technologies in the survey response area were CATI and CAPI (computer assisted telephone interviewing and personal interviewing respectively). In reference to this matrix, these technologies are tools utilized by interviewers. They automate the flow through the questionnaire and provide consistency checks on responses. Thus CATI and CAPII clearly belong in the bottom left cell, reserved for technologies that are tools for use by interviewers to enhance the quality of survey responses. By assisting with the flow through complex questionnaires, these tools may also make it easier for respondents to respond. Therefore, CATI and CAPI also appear in the top left cell. These important technologies simultaneously serve as tools for use by the survey designer and the survey administrator. For example, most CATI systems have a call scheduler who can be utilized by the survey administrator or field director to schedule or reschedule an interview at a convenient time for the respondent, thus easing the burden of responding.

As we look at this matrix, there are tools for interviewer use, respondent use, and for use by the survey designer or administrator. There are tools that help ease response and tools that enhance the quality of responses. However, there are gaps in the matrix cells showing technologies designed to help communicate the reason why a respondent should cooperate. Perhaps the single greatest factor in getting individuals to respond on surveys is to give them sufficient reason (from the respondent's perspective) to do so. It is apparent that our development of technology has generally ignored this important area. Thus this discussant highly recommends future efforts be directed along these lines.

I wish specifically to point out that one of the two papers in this session directly addresses a portion of this area. The paper authored by Link, Armsby, Hubal and Guinn discuss technology that enhances the survey administers' ability to train interviewers on how to avoid refusals during the first 30 seconds of a telephone interview. Thus, it is a tool for survey administrators that helps communicate to potential respondents why they should complete the survey. The other paper in this session looks at Web collection (a tool for respondents) and how that tool enhances the quality of response.

With that overview, we next look more closely at both papers. The Chang and Krosnick paper compares random digit dial (RDD) telephone surveys with two different approaches to Internet surveys. The paper provide results from both a field study and a laboratory study, utilizing surveys of individuals generally focused on political opinions and activities. The paper addresses two areas of concern: the representativeness of the responding sample and the quality of the responses supplied. This was an excellent paper. It is very useful for the survey community to have some work with actual measures of these qualities so that somewhat objective comparisons can be made, particularly concerning Internet surveys.

The paper begins with an excellent discussion of the potential differences between the three modes of collection, and is worth reading for this alone. To the long list of items discussed under response quality, one might add an item concerning the time intervals available for response. There may be a difference in how suitable the different modes are when responses are needed within a very tight time interval. Web collection provides the respondent with more flexibility in terms of when to respond, while a telephone calls pressures the respondent to "do it now". Krosnick indicates that this does not appear to be an issue for the types of surveys addressed in his paper. However, it may affect some Federal agencies' data collections. In trying to publish time sensitive reports, we in NASS often have a window of only a few days for survey response.

The authors provide a wealth of detailed results in comparing different components of quality. For example, they conclude that the Knowledge Network (KN) samples are comparable to the RDD samples in terms of demographic representation, and when weighted all three samples were fairly close. One interesting result concerns measure of non-differentiation. (*Respondent's lack of differentiation among questions with similar response scales is indicative of mindless responding patterns.*) KN respondents receive WebTV equipment in exchange for their participation in surveys, and they showed the greatest amount of non-differentiation. This raises the issue of the use of response incentives in a more general context. Do they lead to quality issues such as non-differentiation? Clearly more research in this area is needed.

Finally, I agree with the authors concluding comments that it is important to compare these different modes of collection in the framework of how they can be complements in future mixed mode designs.

The Link, Armsby, Hubal, and Guinn paper discusses the use of responsive virtual human technology (RVHT) as a tool for teaching refusal avoidance skills to telephone interviewers. The paper addresses some early work in this area. As mentioned earlier, this discussant compliments the authors for working on issues which help fill the “gap” in technology aimed at helping communicate the reasons to respond.

The technology simulates reactions of respondents during the first 30 seconds of a telephone call, and allows interviewers to practice their own reactions and responses. The RVHT tools allow more repetitions for lower cost of this important part of interviewer training. The authors report on early attempts which did not provide completely realistic simulations but which provide optimism for future progress. The simulations ran more slowly than reality, which caused some complaints by experienced interviewers. The real benefit of this type of training is likely to fall to inexperienced interviewers. Similarly the concern with the “slower than life” simulation may give new interviewers false expectations.

This discussant looks forward to following future work on this technology. It will be useful to see results using inexperienced interviewers in a split experiment which compares their subsequent response rates.

I thank the authors, organizer and audience for a very interesting session on technology.

Capitalizing on Technology to Enhance Survey Reporting Discussion

Alan R. Tupek
US Census Bureau

I would like to commend the authors for their thoughtful papers and their innovative work in advancing the use of technology in survey methods. We have come a long way in applying technology to survey methods. The innovative work presented in these papers illustrates just how far we have come.

I will organize my remarks into three broad themes --

I. Innovations, especially innovations that involve process automation, often lead to discoveries that weren't envisioned at the outset.

II. Since this is a conference sponsored by the Federal Committee on Statistical Methodology, what should the role of the Federal Government be as it relates to these projects?

III. Looking ahead. What might be on the horizon that's fueled by these endeavors?

I'll begin with surprises --

I. Innovations, especially innovations that involve process automation, often lead to discoveries that weren't envisioned at the outset.

CATI and CAPI did not reduce response error in surveys, as least as we traditionally measure it through reinterview surveys. In the CPS, CATI and CAPI had little effect on the month-in-sample differences. Computer-assisted methods did not make robots out of interviewers. What it did do was to change the complexity of the survey instruments so that concepts could be measured more precisely. CATI and CAPI also did not reduce costs, but provided the potential to improve collection methods and validation of results.

So what surprises might there be for the paper presented by Polly Armsby and Michael Link? The innovation that they describe is the use of a wide-range of technologies, they call "responsive virtual human technology," to train interviewers to handle the first few seconds of the interview process. While this is truly revolutionary in many respects, a significant side benefit is a learning process about refusal conversions that would not otherwise have occurred. The 48 scenarios discussed in the paper are just the start. Is there an optimum response for each of the 48 scenarios? Or, maybe there are a few preferred responses that work across many of the scenarios. It's telling that interviewers are reacting differently to different verbal cues, yet interviewers ignore the mood and sex of the virtual respondent. Should the interviewers pay more attention to these factors?

In the elections the other day, many voters had to deal with new voting equipment. Here in Montgomery County, we went from punch cards with hanging chads and having to make sure the punch card was properly seated in its place, to touch screen machines. The "improvement" resulted in long lines, probably due to unfamiliarity with the new machines, plus fewer machines compared to the punch card system. Some machines in Montgomery County were misprogrammed. This happened with the new machines in Florida too. I'm sure no one is surprised to hear that. I can't help but think about the potential for programming errors with the touch screen machines. At least with the punch cards, you still have the cards to fall back on as we saw in Florida two years ago. Did the voting officials insist on level one programming? Was there independent programming of all systems?

Let's move on to Internet data collection. The Census Bureau has experimented with Internet data collection. It was available for the short-form in the decennial census. In the American Community Survey, we conducted a split panel test of Internet collection. The households selected for the Internet panel were given the option of responding by Internet but were also sent a questionnaire that they could mail back. The surprise in this test is that the combined mailback and Internet response rate was lower in the Internet panel than the mailback rate in the control panel. Not only did we only get a handful of responses by Internet – around 5% – but it actually reduced the mailback response rate significantly. There are a number of theories as to why this happened. And, we plan to investigate them in future tests of Internet data collection.

When we began developing one-stop shopping for federal statistics, we hadn't envisioned the research program that would evolve. We thought of FedStats as a portal, though we didn't call it that – we called it a window to statistics available on the federal agency websites. Now, through the leadership of Valerie Gregg and Marshall DeBerry, FedStats is a conduit to improving the statistical literacy of the nation.

Let me move on to my second topic.

II. Since this is a conference sponsored by the Federal Committee on Statistical Methodology, what should the role of the Federal Government be as it relates to these projects?

Does it make sense that the government is not at the table?
What should the role of the government be on the use of new technologies?
Should it fund the types of projects we've heard about today?

Regarding the paper presented by Polly Armsby and Michael Link, maybe the question should be "is the survey methods community rich enough to take this highly technical research that is being applied to refusal conversion from its infancy to maturity?" Or, should we let the Defense Department move the technical aspects of this research along to the point that the incremental research cost for the survey community is minimal? The Census Bureau was given the opportunity to participate in this endeavor and we decided to pass. I was one of the ones at the table. We understood the potential, but decided it was too risky. We didn't think the technology was there yet. Let some other research community move this along and when it's mature, we'll take advantage of it. I don't know the right answer. The Census Bureau needs to take risks. It needs to find the right balance between risky long-term projects, short-term research, and production activities. As we've heard from the authors, a lot of progress has been made in the year or so since the Census Bureau was given the opportunity to participate. More progress than I would have expected. If I knew then what I know now, I might have voted differently.

The vital national surveys conducted by the Federal Government, like the Current Population Survey, the National Health Interview Survey, the Consumer Expenditure Survey, the National Crime Victimization Survey, and the American Housing Survey are not likely to embrace the Internet as the sole mode of data collection. Just as the Federal Government uses CATI as one mode in a multi-mode approach for several of these surveys, the Internet will also be used in this way. For similar reasons, the Census Bureau has not embraced RDD surveys except in special situations. The Census Bureau puts a great deal of resources in developing sampling frames that cover the entire target population. In addition, many of its household surveys achieve response rates of over 90 percent. If the Census Bureau did not set and achieve these goals for the vital national surveys, then where will the gold standard come from? How would the other survey organizations know how to weight their survey results?

It's heartening to know that representative samples do make a difference. In the paper by Jon Krosnick, the self-selected samples from the Harris Interactive Surveys provided skewed results even when weighted to detailed characteristics from the Current Population Survey.

III. Looking ahead. What might be on the horizon that's fueled by these endeavors?

I think there's something to be learned from the mandated Internet training instruments that seem to be springing up. For example, I am required to learn about the rules and procedures for using my government-issued credit card for travel. If I don't complete an online training course by such and such a date I must forfeit my credit card. You can skim the pages on the online course as quickly as you can move your mouse and click to the next screen. However, in the end, there's a quiz that requires you to answer most of the questions correctly. Otherwise, you don't pass and you must repeat the exercise.

The Knowledge Network practice of providing a WebTV box and service is something to build on. I'd call this a creative use of incentives. Unfortunately, it's costly, but the good thing is that the cost is spread over a lot of surveys.

There's no end to what one might imagine as uses for the Responsive Virtual Human Technology (RVHT). The training possibilities are endless. But, how can it be used in the interviewing process itself, especially over the web? How might it help with language difficulties, illiteracy, or persons with disabilities? Maybe RVHT can be a virtual boss? "What's the status of that project," delivered in either a sad, had, glad, or mad tone of voice. The human boss needs only to follow-up on the sheepish responses.

In conclusion, I want to thank the presenters and their co-authors for their groundbreaking work and for a stimulating discussion on improving survey methods