Interactive Voice Response (IVR) Case Study: Testing Your Telephone-Based E-Commerce Support

Eric Schaffer, Ph.D., CUA, CPE, is Founder and CEO of Human Factors International, Inc. (HFI). He teaches, consults, and speaks on corporate and governmental Web and GUI interface design issues.

John Sorflaten, Ph.D., CUA, CPE, teaches and consults as a Project Director at HFI. With Eric, he initiated a usability curriculum at a local university in his home town of Fairfield, IA.

Your IVR Insurance Package

E-commerce Depends on IVR E-commerce means commerce supported by electronic technology. This includes the ubiquitous telephone. How can we exploit its convenience, low-cost, and user familiarity for effective commercial applications?

Many organizations seek to reduce the burden of costly customer service staff by using telephone automation. For example, upon calling a credit card support number, the customer hears a recorded request, such as "Please enter your account number." The customer enters the digits using the telephone keypad. The recorded voice then offers options like, "To learn your current balance, press 1. To hear the last 10 transactions, press 2." and so forth. Last in the list of options, you hear "To speak to a customer representative, please press 0."

The bane of such IVR (Interactive Voice Response) systems occurs when the caller presses the "0" button. Because when the customer service representative (CSR) speaks with the caller, they start the "cost-meter." Each second carries a significant loaded labor rate. Some of our corporate customers have indicated that every second they save on the average length of a call means a saving of $120,000 a year. Where CSR employees number in the thousands, savings per second have been reported as a million dollars a year! (Yes, we were surprised, too.)

Given these incentives, organizations seek every opportunity to enhance the efficiency of their CSR and IVR systems. The CSR can work more efficiently when managers insure their computer interface incorporates ergonomic or human factors principles of design (Schaffer, 1998). The IVR can work more efficiently – and keep customers from pressing the dreaded "0" button – by adoption of ergonomic design, as well. Note that the IVR is the first line of "expense defense." Failure of the IVR to assist the caller starts the clock for the more considerable CSR expense. Can we insure against such failure?

Low-cost Insurance Premiums If you, a risk-averse manager, had a chance to purchase insurance that guaranteed the efficacy of your IVR, what would you pay? Would you spend a "person-week" on it? Or even two or three such weeks for a large and sophisticated system? Peanuts, you might reply, when thinking of the person years you have already committed.

You invest in such insurance when you include "usability testing" in your IVR development cycle. Furthermore, the earlier you test, the less costly the insurance. This article will show you what's involved and even empower your team to try it themselves. But first, let's examine alternatives you may have already tried.

Alternatives to Early Testing Some IVR consultants make an interesting suggestion, based on the fact that frustrated callers will at some point quit the IVR protocol, call back, and then select the "rotary telephone" option given at the top of the your IVR menu. This option immediately connects the caller with the CSR because, presumably, the telephone lacks push-button tone technology. However, it has now become an expensive "bypass" operation, neutralizing your IVR investment and accelerating your CSR costs. Obviously, you want to monitor this symptom. By using caller ID your IVR team can see who is dialing back. If they see such repeat calls, then they know your IVR has obstacles. If you have 800 or 900 access, your team can do similar analysis via the ANI.

An additional strategy would be to record data on each call. On the one hand, your team could audiotape all calls (with warning to the caller that calls are monitored for quality control purposes). Your team would analyze the non-completed calls and find problems based on the sequence of IVR prompts, the tones of the entered options, and even uncomplimentary comments uttered by callers. On the other hand, your team could use software to log the number of times a given caller repeated a menu prompt without making a selection. Log other actions such as the menu item on which callers hang up, the length of pauses (i.e., confusing options), and repetitions from the same caller ID or ANI.

In all the above cases, note that you must have already constructed the IVR system to learn where the problems exist. In truth, this isn't insurance. And until you've made changes, it's pure risk exposure. Furthermore, system re-work costs far more than design testing.

Types of Testing If the idea of testing has captured your attention, let's discuss the degree of "fidelity" of such testing. Audiophiles and television home theatre consumers hotly debate issues of lo-fi and hi-fi equipment. Likewise, questions of low-fidelity testing vs. high-fidelity testing intrigues usability experts, as well. Several IVR researchers at GTE Laboratories, Inc. investigated the pros and cons of high-fidelity prototyping of an IVR system (Virzi, Sokolov, Karis, 1996). Their report provides instructive inspiration for low fidelity testing and your consequent low-cost insurance program!

Their high fidelity version of an IVR system included seven databases, over 500 sound files, and took two months to develop. It was an operational IVR system using an Apple Macintosh, but a mockup nonetheless. It used digitally recorded speech, and allowed caller identification, call screening, and automatic callback. The low fidelity version used the "Wizard of Oz" technique in which one experimenter played the part of the computer, reading aloud what the computer would say. Subjects sat in the same room, and indicated which button they pressed in response to the statements. Ten college-age subjects performed the hifi tests, and another 10 did the lo-fi. All subjects received instruction to "think out loud" during the interaction. The sessions were videotaped.

We've already given you the punch line: the low fidelity test performed as well as the high fidelity test. In fact, the researchers recommended "we would not have spent the time and effort to build a high-fidelity prototype" if their only goal was usability testing. (The project did have other goals.) "In fact this is how we currently design

IVR systems in practice." To encourage you in future analyses, here are some of the measures that showed group equivalency. First, the experimenters identified 21 problems with the IVR interface. Comparing the two groups, they found no differences in…

- of problems the subjects uncovered – the hi-fi group found 19 problems, the lo-fi found 20.
- sensitivity of the tests – the number of subjects locating each problem was about the same for each group
- severity of the problems – "eye ball" examination revealed no striking differences in the ability of the lo-fi group to uncover severe problems (both groups uncovered nearly all the problems)

The authors list issues for which a high fidelity prototype can be useful. However, mockups limited to specific questions could serve, as well. A prototype or mockup can test…

- intelligibility of the selected voice or speech synthesis
- concatenation of prompts by the caller
- time to complete menu selections and other performance measures
- display characteristics (e.g., font, images, colors)
- marketing personnel reactions (they like to see verisimilitude – the real thing!)

Furthermore, investment in a prototype can enhance…

- demonstrations for marketing purposes
- uncovering specifications that may not otherwise be obvious
- review of features and functions for documentation and training design purposes

How to Conduct an IVR Usability Test

Introduction The following steps and data represent a demonstration project that Human Factors International, Inc. accomplished on an IVR that served a telecommunications firm that we will call Phones-R-Us. Expert review indicated significant potential for user confusion and consequent overload of the CSR staff. Subjects came from a university population – students and staff. Use the following steps for your tests. Remember the "Wizard of Oz" technique given above – you don't need an operational IVR, although in this test we used one. Here's an overview of how you could present your findings.

IVR Usability Test

- Subjects
- Tasks
- Performance Results
- Satisfaction Results
- Next Steps

Step 1. Get Subjects Choose the number of subjects to match the expected probability of finding a given problem. Big problems need fewer subjects. Subtle problems need more subjects. Experience indicates 10-20 subjects would provide insight into the problems that we anticipated. Our intern tested 16 subjects with telephone experience and varied educational background and gender. He used 2 of the sessions to learn to write the subject's comments rapidly and concisely. We used data from the following 14 subjects. Our intern videotaped five of the interviews in case we wanted to demo the process.

If needed, provide training to give your subjects the same expertise your actual users have. (If you expect a specific background, then recruit – and pay – subjects from your user population.) In our case, we only needed experience using a telephone and age enough to qualify for a telephone card. Here's a subject selection summary:

Subjects

- 14 subjects from a university setting.

- Represents cross-section of US population
- 4 Female (29%)
- 4 English as a second language (ESL)
- Ages 41-50:4; 31-40:1; 21-31:7; 17-20:2
- PhD:1; MA:2; BA:4; HS:7
- Homogeneous: ESL subjects had similar satisfaction ratings as English (5 NSDs)

Comment During data analysis (see below) we wanted to see if ESL made a difference in how subjects felt about the IVR menu. Therefore, we used a statistical test to check for differences between the average scores on each of the 5 satisfaction ratings (given below). "NSD" means No Significant Difference would be found 19 times out of 20 similar tests (the so-called "95% confidence" rating). We used the t-test for unequal variances found in Microsoft Excel. You don't need such confirmation if your own group of test subjects has no particular differentiating characteristic.

Step 2. Determine the Tasks and Test Script

Based on preliminary expert review, we had specific issues we wanted to test. Were our suspicions correct? What percentage of average users would have difficulty? One of us devised 10 test scenarios to meet these needs. Concrete language and specific instances make the test more valid. If necessary, provide any paper documents that would normally be used, such as a credit card statement

On the next page is our test script, with the task scenarios. The data supervisor read the script to maintain consistency of expectation and motivation among subjects.

Step 3. Collect Performance Data Our intern spent about an hour with each subject. He recorded the demographic data (indicated above), then administered the test script. He used a speakerphone so that he could hear the IVR prompts. (Remember, if this were a "low-fidelity" test he would have read out the prompts himself.) He asked the subjects to tell him which button they pressed. He made a point to record the button presses in sequence for each test question. He also recorded his observations and useful subject comments for each button press. At the end of each task, he asked them to describe their experience and degree of difficulty.

Note that a subject may have felt they completed the task correctly because they got a CSR – whether by accident or on purpose. In reality, they were scored as fail because they didn't follow the intent of the design. Subjects had no difficulty with the presence of videotape equipment and its operation. Below, we display a chart of the performance results for easy comprehension. The chart requires tabulation of results (see analysis, below).

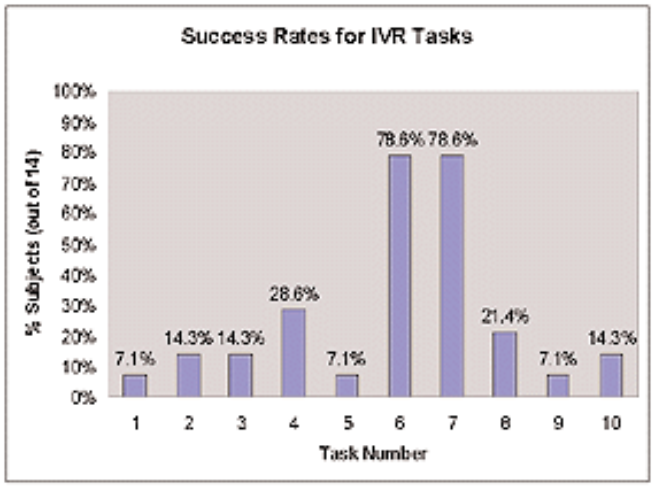

Figure 1

Below, we display subject comments for the two tasks failed by all participants but one. Typically, subjects will not be clear on the source of their confusion. The designer or usability specialist must evaluate each comment for its design significance. Occasionally, a subject offered no comment. We only present two examples here.

TEST SCRIPT

Introduction

Hello,

We are conducting a test of Phone-R-Us's interactive voice recognition system. Your input is valuable to us, and we would appreciate if you could give us feedback as you work your way through the tasks.

This is not a test of your skills, but more importantly, a test of the system. Please be honest with your feedback, as it will only help us in our endeavor to improve this system.

If there are thoughts you have as you are working through the tasks, kindly voice them aloud, so the tester can ask you for some of your thoughts.

Our goal is to obtain feedback on the voice, task flow, and wording of the system prompts. The tester will ask you certain questions at the end regarding your overall impression.

Thanks again! Have fun!

Task List

1. Change your PIN on your calling card.
2. Review the different fees and surcharges to your account as per the Federal Telecommunications Act.
3. Find out information on how you can sign up for Phones-R-Us and Airmiles together.
4. Identify the country code for Austria.
5. You find you have a problem with your account, suspecting that someone is misusing your number (overbilled by $422.10). You would like to speak to a customer service representative immediately to rectify the situation.
6. Check your current rates on your home telephone number.
7. You want to add a personal 800 number to your service.
8. Obtain instructions on how to use your calling card when making a call from France to Japan. Find out the necessary information (such as the country codes to make the call).
9. Credit your account for a call that was incorrectly dialed by an operator (number was supposed to be 770-442-8282, but you dialed 770-442-8228)
10. What number would you dial within the system to link directly to a customer service representative?

Closing Questions

1. What was your overall impression?
2. Did you find the voice pleasing?
3. Were you able to find your way through the system easily?
4. Was the order of prompts easy to understand?
5. Was any wording very difficult to understand?
6. What did you find most difficult?
7. What was the most inviting aspect of this system?
8. Other thoughts?

Thank you.

13 Failed to "Change PIN" (Task 1 – Each bullet is from a different fail subject.)

- I want to update my card, not a service.
- Lots of options and you would need it in writing. I'm not sure if the menus are organized in related groups.
- Use fewer words in sentences.
- Never found words I could recognize as a similar term.
- Nothing was clear. Too many options per number on the menu.
- PIN option wasn't there. I was reaching my limits of patience.
- Was a little confusing. Voice is overly sweet.
- He says too much for each option.
- They assume you know what the words talk about.
- There was nothing specific for calling cards. I got lost.
- All the talking seems to slur together.
- I don't know where to change PIN.

13 Failed to "Review Fees" (Task 2 – Each bullet is from a different fail subject.)

- Was easy to do.
- Lots of choices that had nothing to do with what I wanted to do.
- Would be nice to have it (CSR option) on the main menu, but I can see why they don't. Not so satisfied, a little frustrating.
- Got a little lost in messages
- This was easy task because I am becoming familiar
- Felt satisfied with speed of getting an operator
- Going through all options is a waste if I get a cs rep anyway.
- Opening summary paragraph is annoying. It isn't helpful.
- I feel a … frustration which is not the kind a business wants.
- What's the point of the first menu ... you can't make any selection I wanted a CS rep option on the first menu ...
- From the Phones-R-Us side, it's OK that it took so long, but I get frustrated.

Recall that we also collected statements regarding subjects "overall" impression. Note in the following, that 6 subjects gave favorable statements. They represent 44% of the subjects. Clearly, subjective impressions can be misleading even in the face of severe usability problems.

6 Positive Post-test "Overall Impressions" (Each bullet is from a different positive subject.)

- The system is quite well designed
- It's good
- Relatively usable menu that doesn't waste too much time
- Needs work. Not bad though
- OK, but needs more improvement
- Professional corporate attempt to address a very complicated customer service need – showed much work and effort, yet still room for improvement

8 Negative Post-test "Overall Impressions" (Each bullet is from a different negative subject.)

- A little confusion – a lot offered, but not easily accessible
- Irrelevant, I always ended up going to Customer Service, yet at the beginning there was no option for Customer Service. The whole system couldn't help on half the problems. I could just ask customer service at the beginning and save time.
- A bit too much information on the first menu.
- Could there be a simpler way to get started?
- Slow. They're more interested in mileage than

phone service. Structure not completely clear to find what is needed.

- Complicated because main menu difficult to understand
- Took too long to get anywhere
- Frustration and dis-ease. Feeling of confusion and too much info at each stage.
- Very messy menus. Unclear and confusing

Step 4. Collect Satisfaction Data After the test protocol, each subject filled out a satisfaction questionnaire. (See below.) Because the subject had just attempted 10 tasks, they could easily

General Usability Questionnaire

Note: The information you provide is kept completely confidential and no information is stored on computer media that could identify you as a person.

We would like to get your general impressions of the usability of the telephone interaction that you used today. Please take a few minutes to answer the questions. Thanks for your help in understanding telephone interaction.

Please indicate your agreement or disagreement with the statements in the left column by circling the appropriate number in the right column. If you are undecided or the question appears irrelevant, then circle the middle number (4). (Circle one rating number per statement).

	Disagree						Agree
1. This telephone menu was easy to learn.	1	2	3	4	5	6	7
2. Finding the options to complete the tasks was easy.	1	2	3	4	5	6	7
3. I feel in command of this telephone menu when I am using it.	1	2	3	4	5	6	7
4. There are too many steps required to get something to work.	1	2	3	4	5	6	7
5. Working with this telephone menu is satisfying.	1	2	3	4	5	6	7
6. The terminology is inconsistent with the terminology I know.	1	2	3	4	5	6	7
7. The telephone menu occasionally behaves in a way that can't be understood.	1	2	3	4	5	6	7
8. It is easy to make the telephone menu do exactly what I want.	1	2	3	4	5	6	7
9. The telephone menu seems to disrupt the way I normally do my work.	1	2	3	4	5	6	7
10. The instructions and prompts are helpful.	1	2	3	4	5	6	7
11. It was easy to find information I needed in this telephone menu.	1	2	3	4	5	6	7
12. Learning how to use this telephone menu was difficult.	1	2	3	4	5	6	7
13. There is never enough information when it's needed.	1	2	3	4	5	6	7
14. The telephone menu has a very attractive presentation.	1	2	3	4	5	6	7
15. Error messages are not adequate.	1	2	3	4	5	6	7
16. I have to look for assistance most times when I use this telephone menu.	1	2	3	4	5	6	7
17. These telephone menu features will make me more successful getting what I need.	1	2	3	4	5	6	7
18. I will never learn to use all that is offered in this telephone menu.	1	2	3	4	5	6	7
19. This telephone menu is really very awkward.	1	2	3	4	5	6	7
20. There is too much to learn before one can use this telephone menu.	1	2	3	4	5	6	7
21. It is easy to hear what the options are at each stage.	1	2	3	4	5	6	7

reflect on their subjective reactions. The questionnaire represents five categories of satisfaction (discussed below). We altered the original questionnaire (from other sources) to accommodate IVR technology.

Step 5. Analyze Performance Data In this case, our goal is to show data indicating the extent of the IVR problems. The results guide whether to make design changes or not. Changes themselves presume expert knowledge of IVR design. We represent this phase of the test with the following summary.

Tasks

1. Change PIN
2. Review fees per FTA
3. Learn Phone-R-Us & Airmiles
4. Get code for Austria
5. Check overbilling
6. Check home rates
7. Add personal 800#
8. Learn France to Japan codes
9. Get credit for operator error
10. Get CSR

Overall Results

- 10 tasks X's 14 subjects = 140 test events
- 10 items together averaged 27.1% passing rate
- 8 worst items averaged 11.6% passing rate
- 40% of all failures occur on the first menu item

We scored any task as "fail" if the subject used a different IVR option than what the designer intended. Often, a subject would get a CSR, thinking it was a planned event. While the subject felt a positive outcome, the IVR had failed. We also logged the level of menu at which the subject failed to press the correct phone button. Notably, 40% of all the failures occurred on the first menu item – a prime target for improvement.

We indicate that 140 "test events" constitute the body of evidence. In our data analysis we learned our subjects only passed 27.1% – 38 of the 140 events. Since 22 of those 38 passes occurred in only 2 of the test items, we restated the results. We point out that the 8 worst items merited only 11.6% passing rate – only 16 out of 112 test events. Pretty expensive. The CSR staff has to work hard to keep up with the callers asking for human help.

Step 6. Analyze Satisfaction Data We grouped and averaged data from the satisfaction questionnaire as follows. Where a question implied a negative response, the answer was mathematically converted to match the meaning of the group description. If you use the questionnaire, calculate the mean of the following questions for each category.

a. Learnability: 1, 6, 12, 8, 20 (high priority)

b. Efficiency: 2, 4, 9, 11, 17, 21 (high priority)

c. Control: 3, 7, 8 (low priority)

d. Helpfulness: 10, 13, 15, 16 (low priority)

e. Likability: 5, 14, 19 (medium priority)

Global Metric: .3a + .3b + .1c + .1d + .2e

We established a yardstick of positive merit based on the nature of the scale. Recall that the subjects selected a number from 1 to 7, with 4 representing a neutral point. The next interval above the neutral 4 is 5. Therefore, we set a score of 5 or above as indicating "positive" rating. (See Figure 2)

Comment Two subjects exhibited "experimenter demand" effects-they gave extraordinarily high ratings for many of the questions, although each passed only 4 of 10 tasks. Their global ratings were quite high at 6.6 and 5.9 out of a possible 7. We investigated any demographic correlate that would motive the high rating – such as English language shortcomings, and consequent misunder-

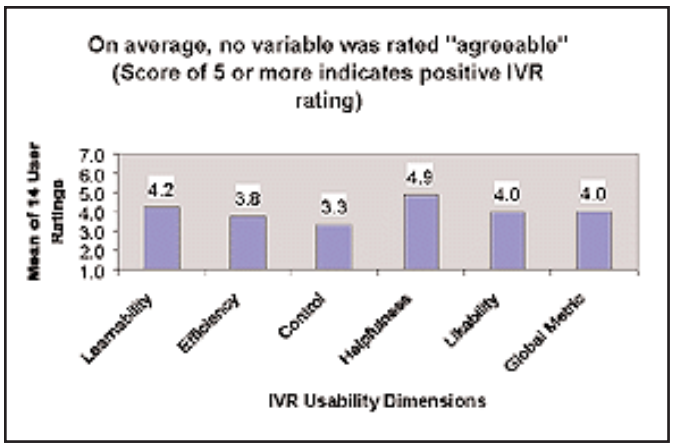

Figure 2

standing of the situation. While one of the subjects indeed was "ESL" (English as a second language), the other subject spoke English as a first language. We mention this aspect of our data collection to illustrate issues that may arise in your own testing. "Experimenter demand" effect is a known phenomenon of psychological testing in which the subject attempts to give responses to questions they feel the experimenter wants, regardless of their personal experience. While we cannot say for sure this is the case, we had data from enough other subjects to establish a reasonable range of response patterns.

Step 7. Next Steps

Introduction Examination of subject comments reveal these sources of confusion …

- menus that were too lengthy for easy comprehension
- unclear vocabulary terms
- inconsistent terminology
- organization of the menus
- unclear categories
- too many steps (menus too deep)

Furthermore, in 40% of the 120 events failure occurred on the first menu, while slightly fewer fails occurred across the other 4 levels of menu. The immediate presentation of difficulties – at the first menu – indicates a need to focus on general design principles including category definition and minimizing demand on the caller's short-term memory.

Sources and Types of Design Help The human factors literature provides excellent guidance on issues of IVR design. Much work has been accomplished by Vrizi and Resnick in the context of their research for GTE Laboratories. See their articles listed in the bibliography. In particular, they report a technique of "skip and scan" that we have found very useful when a caller is faced with six to 15 menu options. They have also conducted usability studies comparing skip and scan to a more conventional method. They found that skip and scan was superior for both novice and expert users in all but the first few trials. They used 36 tasks and two IVR applications for the tests.

Menu Level	# of Failed Events	% Failure
1	56	40.0%
2	9	6.4%
3	25	17.9%
4	8	5.7%
5	4	2.9%
Total	102	72.9%

Some managers may question the value of IVR for their business. A study of public attitudes about IVR as well as telephone answering machines indicate that attitudes varied strongly by age (Kaatz, Aspden, and Reich, 1977). But more importantly, they found that "information rich" persons were not any more positively inclined to IVR than the "information poor."

The most significant predictor of IVR acceptance was the quality of one's most recent experience

with the technology. This conclusion implies that good IVR design begets positive acceptance.

What do callers want in IVR voice quality? In a study of 84 personality traits, more than 50% of the 50 subjects wanted the following (Chin, 1996):

- Practical (78% of subjects)
- Intelligent (76%)
- Courteous (72%)
- Efficient (68%)
- Straightforward (60%)
- Methodical (54%)
- Sophisticated (50%)

Other top scoring qualities were selected by more than 28% of the subjects:

- Progressive (44%)
- Alert (30%)
- Scientifically minded (30%)
- Imaginative (28%)

To use this data, you could create a "Voice Quality Questionnaire" similar to the "Satisfaction Questionnaire." Use the 11 adjectives listed above, and let your subjects rate sample voice messages.

Standards The benefits of a standard approach to designing IVR can be obvious to managers faced with large-scale development. Such challenges require coordination among numerous developers and consistency between various menus and even projects. The Human Factors and Ergonomics

If the caller's task is...*	Then use a...
Navigating between menus	Selection widget
Selecting one from a list	Pick One
Less than seven items	Scan and Pick One
Between six and 15 items	Enter Alphanumeric / Direct Select or
More than 15 items	Search for Selection
Selecting more than one item from a list	Pick Many
Selecting item(s) from a list by name	Search for Selection
Entering alphanumerics	Enter Alphanumeric/ Direct Select
Transferring shares between accounts or purchasing new shares with funds in current holdings	Exchange
Purchasing shares	Purchase
Redeeming shares	Redeem
Playback of data messages with multiple, addressable parts	Segmented List Playback
Playback of data messages with one item or single message	Nonsegmented Playback
Escaping from current activity	Control/Trap Menu
List management	List Editing
Accessing the voice system	Greeting Menu or Greeting Message

*Thanks to John Ims, PhD, and Gary Dorst, PhD, for creating this table.

Society (HFES) has coordinated development of U.S. standards with the international ISO 9241 human-computer interaction standard. The HFES effort is called ANSI/HFES 200 of which section 9 covers Interactive Voice Response, among other voice design topics (Blanchard, 1997).

Standards can be simple and easy to follow – in fact, they will be rejected by developers without such crafting. On the previous page is a sample list of "call flow structures" that meet the needs of a given organization. Our company, HFI, has found this an excellent approach to designing IVR standards. Each organization requires different templates of such call flow structures that readily guide the development staff. Upon selecting a task, the developer need only attend to the recommended template that describes the method.

Bibliography

Blanchard, H.E. (1997). Standards: HCI standards in the United States. *SIGCHI Bulletin* 29(2)

Chin, J.P. (1996). Personality trait attributions to voice mail user interfaces. *CHI '96 Electronic Proceedings*, Short Papers.

Gardner-Bonneau, D. J. (1992). Human factors problems in interactive voice response (IVR) applications: do we need a guideline/standard? *Proceedings of the Human Factors Society 36th Annual Meeting*, 1, 222-226.

Kaatz, J., Aspden, P, and Reich, W.A. (1997). A national survey of opinions about voice response units and telephone answering machines: three surveys and a framework. *Behavior and Information Technology* 16(3), 125-144.

Resnick, P. and Virzi, R. A. (1992). Skip and scan: cleaning up telephone interfaces. *Proceedings of ACM CHI '92* Conference on Human Factors in Computing Systems, May 3-7, 419-426.

Resnick, P. and Virzi, R.A. (1995). Relief from the audio interface blues: expanding the spectrum of menu, list, and form styles research contributions. *ACM Transactions on Computer-Human Interaction* 2(2), 145-176.

Schaffer, E. M. (1998). How to design effective Graphical User Interfaces – 3 day seminar. Human Factors International, Inc., Fairfield, IA, 52556.

Virzi, R. A., Resnick, P. and Ottens, D (1992). Skip and scan telephone menus: user performance as a function of experience. *Proceedings of the Human Factors Society 36th Annual Meeting*, 1, 211-215.

Virzi, R.A., Sokolov, J.L., Karis, D (1996). Usability problem identification using both low- and high-fidelity prototypes. *CHI '96 Electronic Proceedings*, Papers.